

Narrowest-Over-Threshold Detection of Multiple Change-points and Change-point-like Features

Rafal Baranowski, Yining Chen and Piotr Fryzlewicz*

London School of Economics and Political Science

September 2, 2016

Abstract

We propose a new, generic and flexible methodology for nonparametric function estimation, in which we first estimate the number and locations of any features that may be present in the function, and then estimate the function parametrically between each pair of neighbouring detected features. Examples of features handled by our methodology include change-points in the piecewise-constant signal model, kinks in the piecewise-linear signal model, and other similar irregularities, which we also refer to as generalised change-points.

Our methodology works with only minor modifications across a range of generalised change-point scenarios, and we achieve such a high degree of generality by proposing and using a new multiple generalised change-point detection device, termed Narrowest-Over-Threshold (NOT). The key ingredient of NOT is its focus on the smallest local sections of the data on which the existence of a feature is suspected. Crucially, this adaptive localisation technique prevents NOT from considering subsamples containing two or more features, a key factor that ensures the general applicability of NOT.

For selected scenarios, we show the consistency and near-optimality of NOT in detecting the number and locations of generalised change-points, and discuss how to extend the proof to other settings. The NOT estimators are easy to implement and rapid to compute: the entire threshold-indexed solution path can be computed in close-to-linear time. Importantly, the NOT approach is easy to extend by the user to tailor to their own needs. There is no single competitor, but we show that the performance of NOT matches or surpasses the state of the art in the scenarios tested. Our methodology is implemented in the R package **not**.

Key words: Break-point detection, knots, piecewise polynomials, segmentation, splines, smoothing.

*Address: Department of Statistics, Columbia House, Houghton Street, London, WC2A 2AE, UK
Emails: {r.baranowski, y.chen101, p.fryzlewicz}@lse.ac.uk

1 Introduction

This paper considers the canonical univariate statistical model

$$Y_t = f_t + \varepsilon_t, \quad t = 1, \dots, T, \quad (1.1)$$

where the deterministic and unknown signal f_t is believed to display some regularity across the index t , and the stochastic noise ε_t is exactly or approximately centred at zero. Despite the simplicity of model (1.1), inferring information about f_t remains a task of fundamental importance in modern applied statistics and data science. We now mention a selection of applications in which the task of interest reduces to estimating or making inference on f_t or its functionals. In the analysis of DNA copy number data in genomics, f_t is usually modelled as piecewise-constant and the typical task is to estimate change-points in f_t (Olshen et al., 2004). In mass spectrometry, it is often of interest to detect peaks in f_t (Antoniadis et al., 2010). In applied financial econometrics, a key task is to identify current trends in financial markets, which can be translated into searching for significant recent changes in some characteristics of f_t , e.g. the local slope (Schroeder and Fryzlewicz, 2013). In climatology, detecting changes in the trends of temperature data (Cahill et al., 2015) can also be formalised as estimating changes in the slope of f_t while modelling it as piecewise-linear. In astrophysics, detecting Gamma-Ray Bursts (GRB; Kolaczyk, 1997) typically requires delicate statistical work resulting in the separation of f_t into slowly-varying background and faster-changing GRB's.

Depending on the nature and complexity of the statistical task involving f_t , a wider or narrower range of tools are at the statistician's disposal, and we provide a non-exhaustive list of the main approaches below. When the task is the simple estimation of f_t , linear methods such as kernel smoothing (Wand and Jones, 1994), spline smoothing (De Boor, 2001) or local polynomial regression (Fan and Gijbels, 1996; Simonoff, 2012) typically provide a useful reference point, and robust smoothing techniques (such as median filtering; Koch, 1996) may be of interest if the distribution of ε_t is heavy-tailed. On the other hand, when the interest is in more interpretable estimation, for example in the detection of “features” in f_t such as jumps or kinks, then more involved, non-linear techniques are usually required. If f_t is modelled as piecewise-constant and it is of interest to detect its change-points, several techniques are available, and we only mention a selection of older and more recent approaches. When ε_t is assumed to be Gaussian, both non-penalised and penalised least squares approaches were first considered by Yao and Au (1989). For specific choices of penalty functions, see e.g. Yao (1988) and Lavielle (2005). The Gaussianity assumption on the noise ε_t is relaxed to exponential family distributions in Lee (1997), Hawkins (2001) and Frick et al. (2014). In particular, Frick et al. (2014) also provide confidence intervals for the location of the estimated change-points. Note that often this penalty-type approach requires a computational cost of at least $O(T^2)$, with the exception of the estimator proposed by Killick et al. (2012a), which achieves a linear computational cost (thus called the “Pruned Exact Linear Time”, or PELT), but requires further assumption that change-points are separated by time intervals drawn independently from some probability distribution, a scenario in which considerations of statistical consistency are not generally possible. A nonparametric version of PELT is investigated by Haynes et al. (2016a). Another general approach is based on the idea of Binary Segmentation (BS; Vostrikova, 1981), which can be viewed as a greedy approach with

a limited computational cost. Its popular variants include the circular binary segmentation (CBS; Olshen et al., 2004) and the Wild Binary Segmentation (WBS; Fryzlewicz, 2014). A more complete review in terms of up-to-date publications, software and applications can be found in the online repository *changeoint.info* maintained by Killick et al. (2012b). More general change-point problems, in which f_t is modelled as piecewise-parametric (not necessarily piecewise-constant) between “knots”, the number and locations of which are unknown and need to be estimated, have attracted less interest in the literature and overwhelmingly focus on linear trend detection. Among them, we mention the approach based on least squares principle and Wald-type tests by Bai and Perron (1998), and trend filtering (Tibshirani, 2014; Lin et al., 2016).

The aim of this work is to propose a new, generic approach to the problem of detecting an unknown number of “features” occurring at unknown locations in f_t . By a feature, we mean a characteristic of f_t , occurring at a location t_0 , that is detectable by considering a sufficiently large subsample of data Y_t around t_0 . Examples include: change-points in f_t when it is modelled as piecewise-constant, change-points in the first derivative when f_t is modelled as piecewise-linear and continuous, and discontinuities in f_t when it is modelled as piecewise-linear but without the continuity constraint. We will provide a precise description of the type of features we are interested in later on. Moving beyond f_t only, our approach will also permit the detection of similar features present in some distributional aspects of ε_t , for example in its variance. Since all types of features we consider describe changes in a parametric description of f_t , we use the terms “feature detection” and “change-point detection” interchangeably throughout the paper. Occasionally, for precision, we will be referring to change-point detection in the piecewise-constant model as the “canonical” change-point problem, while our general feature detection problem will sometimes be referred to as a “generalised” change-point problem.

Core to our approach is a particular blend of “global” and “local” treatment of the data Y_t in the search for the multiple features that may be present in f_t , a combination that gives our method a multi-scale character. At the first “global” stage, we randomly draw a number of subsamples $(Y_s, Y_{s+1}, \dots, Y_e)'$, where $1 \leq s < e \leq T$. On each subsample, we assume, possibly erroneously, that *only one* feature is present and use a tailor-made contrast function derived (according to a universal recipe we provide later) from the likelihood theory to find the most likely location of the feature. We retain those subsamples for which the contrast *exceeds a certain user-specified threshold*, and discard the others. Amongst the retained subsamples, we search for the one drawn on the *narrowest* interval, i.e. one for which $e - s$ is the smallest: it is this step that gives rise to the name *Narrowest-Over-Threshold* (NOT) for our methodology. The focus on the narrowest interval constitutes the “local” part of the method, and is a key ingredient of our approach which ensures that with high probability, at most one feature is present in the selected interval. This key observation gives our methodology a general character and allows it to be used, only with minor modifications, in a wide range of scenarios, including those described in the previous paragraph. Having detected the first feature, the algorithm then proceeds recursively to the left and to the right of it, and stops, on any current interval, if no contrasts can be found that exceed the threshold.

Besides its generic character, other benefits of the proposed methodology include low computational complexity, ease of implementation, accuracy in the detection of the feature locations, and the fact that it enables parametric (and hence: interpretable) estimation of

the signal on each section delimited by a pair of neighbouring estimated features. Regarding the computational complexity, the facts that only a limited number of data subsamples, M , need to be drawn (we provide precise bounds later; with finitely many change-points, one can take $M = O(\log T)$ in general), and that typical contrasts are computable in linear time, lead to a computational complexity of $O(MT)$ for the entire procedure. Moreover, the entire threshold-indexed solution path can also be computed efficiently, in typically close-to-linear time, as observed from our numerical experiments. Regarding the estimation accuracy, in the scenarios we consider theoretically, our procedure yields near-optimal rates of convergence for the estimators of feature locations.

Importantly, the flexible character of our methodology leaves it open to possible extensions and modifications. Indeed, borrowing words from Sweldens and Schröder (2000), who advocated “building your own wavelets at home”, we also view our proposal as flexible enough to enable the user to “construct their own feature detector at home”, e.g. by proposing their own specialised contrast functions, or by data-adaptively choosing the most suitable contrast function from a pre-specified dictionary (which would lead to mixed-type feature detection). Although these extensions are not covered in the current work, we view this modularity and flexibility offered by our methodology as an important aspect of our proposal.

On a broader level, our methodology promotes the idea of “fitting simple models on subsets of the data (the local aspect), and then aggregating the results to obtain the overall fit (the global aspect)”, an idea also present in the Wild Binary Segmentation method of Fryzlewicz (2014). However, we emphasise that the way the simple models (here: models containing *at most one* change-point or other feature) are fitted in the NOT and WBS methods are entirely different and have different aims. Unlike the WBS, the NOT methodology focuses on the *narrowest* intervals of the data on which it is possible to locate the feature of interest. It is this focus on the narrowest intervals that enables NOT to extend well beyond mere change-point detection for a piecewise-constant f_t , the latter being the sole focus of the WBS method. The lack of the narrowest-interval focus in the WBS and BS methods means that it is not applicable to more general feature detection, and we explain the mechanics of this phenomenon briefly in the following simple example.

Consider a continuous piecewise-linear signal that has two change-points in its first derivative:

$$f_t = \begin{cases} \frac{1}{350}t, & t = 1, \dots, 350, \\ 1, & t = 351, \dots, 650, \\ \frac{1001}{350} - \frac{1}{350}t, & t = 651, \dots, 1000. \end{cases} \quad (1.2)$$

If we approximate f_t using a piecewise-linear signal with only one change-point in its derivative, then the best approximation (in terms of minimising the ℓ_2 distance) will result in an estimated change-point at $t = 500$, which is away from the true ones at $t = 350$ and $t = 650$, as is illustrated in Figure 1. Therefore, taking the entire sample of data starting at $s = 1$ and ending at $e = 1000$, and searching for one of its multiple change-points by fitting, via least squares, a triangular signal with a single change-point, does not make sense. NOT avoids this issue because of its unique feature of picking the *narrowest* intervals which are likely to contain only one change-point. To understand the mechanics of this key feature, imagine that now f_t is observed with noise. Through its pursuit of the narrowest intervals, NOT will

ensure that, with high probability, some suitably narrow intervals around the change-points $t = 350$ and $t = 650$ are considered. More precisely, by construction, they will be *narrow enough to contain only one change-point each*, but wide enough for the designed contrast (see Section 2.3.2 for more on contrasts) to indicate the existence of the change-point within both of them. The designed contrast function will indicate the right location of the change-point (modulo the estimation error) if only one change-point is present in the data subsample considered, unlike in the situation described earlier in which multiple change-points were included in the chosen interval. More details on this example are presented in Section 3.3.

We note that this example is different from the canonical change-point detection problem (i.e. piecewise-constant signal with multiple change-points), where if we approximate the signal using a piecewise-constant function with only one change-point, the change-point of the fitted signal will always be among the true ones (Venkatraman, 1992). Since the latter property does not hold in most generalised change-point detection problems, this highlights the need for new methods with better localisation of the feature of interest, such as our NOT algorithm. In the final stages of preparing this manuscript, we learned that Fang et al. (2016) independently considered a related shortest-interval idea in the context of the canonical change-point detection problem. However, they did not consider it as a springboard to more general feature detection problems, which is the key motivation behind NOT and its most valuable contribution.

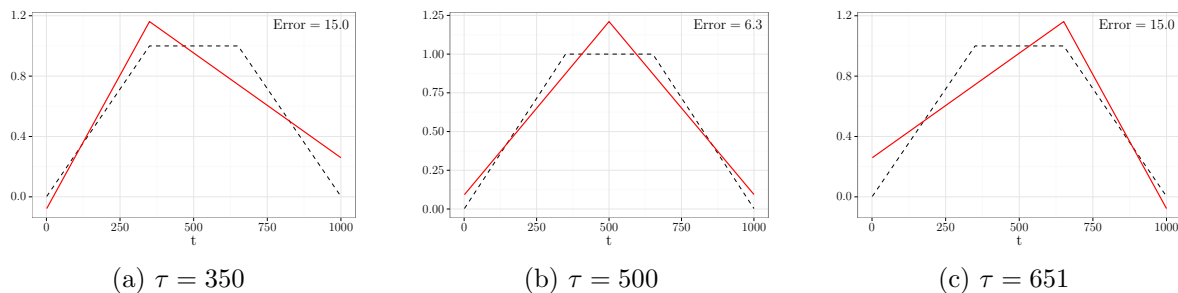


Figure 1: Best ℓ_2 approximation of the true signal (dashed) via a triangular signal with a single change-point, the location of which is fixed at the left change-point (left panel), halfway between the true change-points (middle panel) and at the right change-point (right panel). Approximation errors (in terms of squared ℓ_2 distance) are given in the top-right corners of the corresponding panels.

To summarise, in the NOT approach, we propose a new “modus operandi” in statistical smoothing, by providing a novel, general, flexible framework for feature detection and interpretable signal estimation. The procedure is fast, accurate, easy to code and to extend by the user to tailor to their own needs. Its implementation is provided in the R package **not** (Baranowski et al., 2016b).

The remainder of this paper is organised as follows. In Section 2, we give a more mathematical description of NOT. In particular, we consider NOT in four scenarios, each with a different form of structural change in the mean and/or variance. For the development of both theory and computation, in each scenario, we also introduce the tailor-made contrast function derived from the generalised likelihood ratio (GLR), which is used to detect features

within each subsample. Theoretical properties of NOT, such as its consistency and convergence rates are also provided. Section 3 deals with the computational aspects of NOT, while a comprehensive simulation study is carried out in Section 4, where we compare NOT with the state-of-art change-point detection tools. In Section 5, we consider data examples of oil price, global temperature anomalies and London housing data. All proofs are deferred to the appendices.

2 Methodology

2.1 Setup

To describe the main framework of NOT, we consider a simplified version of (1.1), where $\mathbf{Y} = (Y_1, \dots, Y_T)'$ is modelled through

$$Y_t = f_t + \sigma_t \varepsilon_t, \quad t = 1, \dots, T, \quad (2.1)$$

where f_t is the signal, $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ is the standardised independent and identically distributed (i.i.d.) Gaussian noise, and where σ_t is the noise's standard deviation at time t . We note that the normality assumption facilitates the technical presentation of our results, but the entire framework can be extended to other noise distributions. Numerical examples involving other noise distributions can be found in Section 4.

We assume that (f_t, σ_t) can be partitioned into $q + 1$ segments, with q unknown distinct change-points $0 = \tau_0 < \tau_1 < \dots < \tau_q < \tau_{q+1} = T$. Here the value of q is not pre-specified and can grow with T . For each $j = 1, \dots, q + 1$ and for $t = \tau_{j-1} + 1, \dots, \tau_j$, the structure of (f_t, σ_t) is modelled parametrically by a local (i.e. depending on j) real-valued d -dimensional parameter vector Θ_j (with $\Theta_j \neq \Theta_{j-1}$), where d is known and typically small. In addition, we require the minimum distance between consecutive change-points to be greater than d for the purpose of identifiability. In other words, (f_t, σ_t) can be divided into q different segments, each from the same parametric family of much simpler structure. Even if the main goal is not change-point detection, the class of piecewise-parametric functions is rich enough for function estimation, as any function could be approximated arbitrarily well in L_p ($0 < p < \infty$) by a piecewise-parametric function with enough segments (DeVore, 1998).

Some commonly-encountered scenarios are listed below, where the following holds inside the j -th segment for each $j = 1, \dots, q + 1$:

(S1) **Constant variance, piecewise-constant mean:**

$$\sigma_t = \sigma_0 \text{ and } f_t = \theta_j \text{ for } t = \tau_{j-1} + 1, \dots, \tau_j.$$

(S2) **Constant variance, continuous and piecewise-linear mean:**

$\sigma_t = \sigma_0$ and $f_{\tau_{j-1}+1} = \theta_{j,1}$, $f_t = f_{t-1} + \theta_{j,2}$ for $t = \tau_{j-1} + 2, \dots, \tau_j$, with the additional constraint of

$$\theta_{j,1} + \theta_{j,2}(\tau_j - \tau_{j-1} - 1) = \theta_{j+1,1} - \theta_{j+1,2}$$

for $j = 1, \dots, q$. Therefore, $t \in \{\tau_1, \dots, \tau_q\}$ if and only if $f_{t-1} + f_{t+1} \neq 2f_t$.

(S3) **Constant variance, piecewise-linear (but not necessarily continuous) mean:**

$$\sigma_t = \sigma_0 \text{ and } f_{\tau_{j-1}+1} = \theta_{j,1}, f_t = f_{t-1} + \theta_{j,2} \text{ for } t = \tau_{j-1} + 2, \dots, \tau_j.$$

(S4) **Piecewise-constant variance, piecewise-constant mean:**

$$f_t = \theta_{j,1} \text{ and } \sigma_t = \theta_{j,2} > 0 \text{ for } t = \tau_{j-1} + 1, \dots, \tau_j.$$

Since σ_0 in (S1)–(S3) acts as a nuisance parameter, in the rest of this manuscript, for simplicity we assume that its value is known. If it is unknown, then it can be estimated accurately using the Median Absolute Deviation (MAD) method (Hampel, 1974). More specifically, the MAD estimator of σ_0 is defined as $\hat{\sigma} = \text{Median}\{|Y_2 - Y_1|, \dots, |Y_T - Y_{T-1}|\} / \{\Phi^{-1}(3/4)\sqrt{2}\}$ in Scenario (S1) and as $\hat{\sigma} = \text{Median}\{|Y_1 - 2Y_2 + Y_3|, \dots, |Y_{T-2} - 2Y_{T-1} + Y_T|\} / \{\Phi^{-1}(3/4)\sqrt{6}\}$ in Scenarios (S2) and (S3), where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution.

Both the methodology and the theory developed below can readily be extended to handle more complicated cases in which the signal within the segments is non-linear (e.g. higher-order-polynomial, a case illustrated in Section 4). In all of the above-listed scenarios, we focus on structure changes in the mean or the first two moments in the univariate setting. Nevertheless, our framework can be extended to handle multivariate observations, or other more complex structure changes such as autocovariance in time series. In addition, as mentioned earlier, the normality assumption of the noise can be relaxed as well.

2.2 Main idea

We now describe the main idea of NOT formally. In the first step, instead of directly using the entire data sample, we randomly extract subsamples, i.e. vectors $(Y_s, Y_{s+1}, \dots, Y_e)'$, where s and e are integers drawn (independently with replacement) uniformly from the set $\{1, \dots, T\}$ that satisfy $1 \leq s < e \leq T$ and $e - s > 2(d - 1)$. Let $\ell(Y_s, \dots, Y_e; \Theta)$ be the likelihood of Θ given $(Y_s, \dots, Y_e)'$. We then compute the generalised log-likelihood ratio (GLR) statistic for all potential single change-points within the subsample and pick the maximum, that is,

$$\begin{aligned} \mathcal{R}_{s,e}^b(\mathbf{Y}) &= 2 \log \left[\frac{\sup_{\Theta^1, \Theta^2} \{\ell(Y_s, \dots, Y_b; \Theta^1) \ell(Y_{b+1}, \dots, Y_e; \Theta^2)\}}{\sup_{\Theta} \ell(Y_s, \dots, Y_e; \Theta)} \right]; \\ \mathcal{R}_{s,e}(\mathbf{Y}) &= \max_{b \in \{s+d-1, \dots, e-d\}} \mathcal{R}_{s,e}^b(\mathbf{Y}). \end{aligned} \quad (2.2)$$

If constraints are in place between Θ_j and Θ_{j+1} for any $j = 1, \dots, q$ (e.g. as in (S2)), the supremum in the numerator of (2.2) is taken over the set that only contains elements of form $\Theta^1 \times \Theta^2$ satisfying these constraints. Otherwise, as in (S1), (S3) and (S4), (2.2) can be simplified to

$$\mathcal{R}_{s,e}^b(\mathbf{Y}) = 2 \log \left\{ \frac{\sup_{\Theta} \ell(Y_s, \dots, Y_b; \Theta) \sup_{\Theta} \ell(Y_{b+1}, \dots, Y_e; \Theta)}{\sup_{\Theta} \ell(Y_s, \dots, Y_e; \Theta)} \right\}.$$

The above procedure is repeated on M pairs of randomly drawn integers $(s_1, e_1), \dots, (s_M, e_M)$.

In the second step, we test all the $\mathcal{R}_{s_m, e_m}(\mathbf{Y})$ for $m = 1, \dots, M$ against a given threshold ζ_T , and pick the one corresponding to the interval $[s_m^*, e_m^*]$ that has the smallest length.

Once a change-point is found in $[s_{m^*}, e_{m^*}]$ (i.e. the b^* that maximises $\mathcal{R}_{s_{m^*}, e_{m^*}}^b(\mathbf{Y})$), the same procedure is then repeated recursively to the left and to the right of it, until no further significant GLRs can be found.

After finding all the change-points, one can estimate the signals within each segment using standard methods such as least squares or maximum likelihood. Note that spline regression can be viewed as a multiple change-point detection problem set in the context of polynomial segments that are continuously differentiable but have discontinuous higher order derivatives at the change-points between these segments. From this perspective, one can also think of NOT as an adaptive way of picking the number and the location of knots from the data for the traditional spline regression.

2.3 Log-likelihood ratios and contrast functions

In many applications, the GLR (2.2) in NOT can be simplified with the help of “contrast functions” under the setting of Gaussian noise. More precisely, for every integer triple (s, e, b) with $1 \leq s < e \leq T$, our aim is to find $\mathcal{C}_{s,e}^b(\mathbf{Y})$ such that:

- (a) $\operatorname{argmax}_b \mathcal{C}_{s,e}^b(\mathbf{Y}) = \operatorname{argmax}_b \mathcal{R}_{s,e}^b(\mathbf{Y})$,
- (b) heuristically speaking, the value of $\mathcal{C}_{s,e}^b(\mathbf{Y})$ is relatively small if there is no change-point in $[s, e]$,
- (c) the formulation of $\mathcal{C}_{s,e}^b(\mathbf{Y})$ mainly consists of taking inner products between the data and contrast vectors, which facilitates the development of both computation and theory, particularly if the contrast vectors can be taken to be mutually orthonormal.

In the following, we give the contrast functions corresponding to (S1)–(S4). We note that this approach recovers the CUSUM statistic in (S1), which is popular in this canonical change-point detection setting. One can view the resulting statistics as generalisations of CUSUM to other scenarios.

2.3.1 Scenario (S1)

Here f_t is piecewise-constant. For any integer triple (s, e, b) with $1 \leq s < e \leq T$ and $s \leq b \leq e - 1$, we define the contrast vector $\boldsymbol{\psi}_{s,e}^b = (\psi_{s,e}^b(1), \dots, \psi_{s,e}^b(T))'$ with

$$\psi_{s,e}^b(t) = \begin{cases} \sqrt{\frac{e-b}{l(b-s+1)}}, & t = s, \dots, b \\ -\sqrt{\frac{b-s+1}{l(e-b)}}, & t = b+1, \dots, e \\ 0, & \text{otherwise,} \end{cases} \quad (2.3)$$

where $l = e - s + 1$. Also, if $b \notin \{s, s+1, \dots, e-1\}$, then we set $\psi_{s,e}^b(t) = 0$ for all t . As an illustration, plots of $\boldsymbol{\psi}_{s,e}^b$ with different (s, e, b) are shown in Figure 2(a).

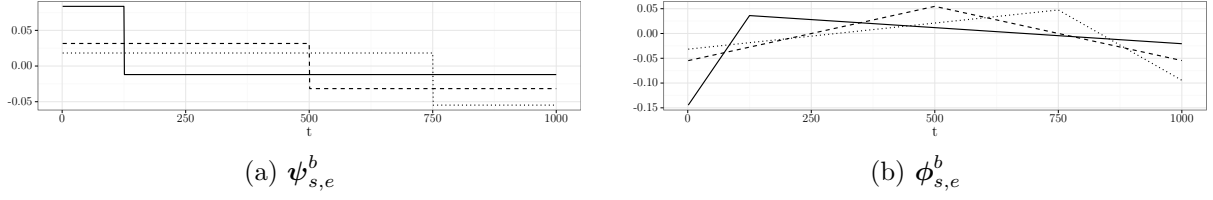


Figure 2: Plots of $\phi_{s,e}^b$ and $\psi_{s,e}^b$ given by, respectively, (2.3) and (2.6) for $s = 1$, $e = 1000$ and several values of b . Solid line: $b = 125$; dotted line: $b = 500$; dashed line: $b = 750$.

For any vector $\mathbf{v} = (v_1, \dots, v_T)'$ we define the contrast function as $\mathcal{C}_{s,e}^b(\mathbf{v}) = \sqrt{\langle \mathbf{v}, \psi_{s,e}^b \rangle^2} = |\langle \mathbf{v}, \psi_{s,e}^b \rangle|$. Therefore, if $s \leq b \leq e - 1$, then

$$\mathcal{C}_{s,e}^b(\mathbf{v}) = \left| \sqrt{\frac{e-b}{l(b-s+1)}} \sum_{t=s}^b v_t - \sqrt{\frac{b-s+1}{l(e-b)}} \sum_{t=b+1}^e v_t \right|. \quad (2.4)$$

Otherwise, $\mathcal{C}_{s,e}^b(\mathbf{v}) = 0$. This recovers the well-known CUSUM statistic in the change-point detection literature. It can be shown that $[\mathcal{C}_{s,e}^b(\mathbf{Y})]^2 = \sigma_0^2 \mathcal{R}_{s,e}^b(\mathbf{Y})$ for every (s, e, b) with $1 \leq s \leq b < e \leq T$, thus $\mathcal{C}_{s,e}^b(\cdot)$ fulfills the aforementioned requirements for the contrast function.

In addition, with a slight abuse of notation, for any $1 \leq s < e \leq T$, we define the constant vector for the interval $[s, e]$ as

$$\mathbf{1}_{s,e}(t) = \begin{cases} (e-s+1)^{-1/2}, & t = s, \dots, e \\ 0, & \text{otherwise} \end{cases}, \quad (2.5)$$

and write $\mathbf{1}_{s,e} = (\mathbf{1}_{s,e}(1), \dots, \mathbf{1}_{s,e}(T))'$. Then it is easy to check that $\mathbf{1}_{s,e}$ and $\psi_{s,e}^b$ are orthonormal. This explains why the CUSUM is invariant to shifts in the mean.

2.3.2 Scenario (S2)

Here f_t is piecewise-linear and continuous. For any triple (s, e, b) with $1 \leq s < e \leq T$ and $s+1 \leq b \leq e-1$, consider the contrast vector $\phi_{s,e}^b = (\phi_{s,e}^b(1), \dots, \phi_{s,e}^b(T))'$ with

$$\phi_{s,e}^b(t) = \begin{cases} \alpha_{s,e}^b \beta_{s,e}^b \left[\{3(b-s+1) + (e-b) - 1\}t - \{b(e-s) + 2s(b-s+1)\} \right], & t = s, \dots, b \\ -\frac{\alpha_{s,e}^b}{\beta_{s,e}^b} \left[\{3(e-b) + (b-s+1) + 1\}t - \{b(e-s) + 2e(e-b+1)\} \right], & t = b+1, \dots, e \\ 0, & \text{otherwise.} \end{cases} \quad (2.6)$$

where $\alpha_{s,e}^b = \left(\frac{6}{l(l^2-1)(1+(e-b+1)(b-s+1)+(e-b)(b-s))} \right)^{1/2}$, $\beta_{s,e}^b = \left(\frac{(e-b+1)(e-b)}{(b-s)(b-s+1)} \right)^{1/2}$ and $l = e - s + 1$. If $b \notin \{s+1, \dots, e-1\}$, then we set $\phi_{s,e}^b(t) = 0$ for all t . We illustrate the structure of $\phi_{s,e}^b$ in Figure 2(b). The contrast function is then defined as

$$\mathcal{C}_{s,e}^b(\mathbf{v}) = \sqrt{\langle \mathbf{v}, \phi_{s,e}^b \rangle^2} = |\langle \mathbf{v}, \phi_{s,e}^b \rangle|, \quad (2.7)$$

To explain the rationale behind $\phi_{s,e}^b$, we first define the “linear” vector for the interval $[s, e]$, $\gamma_{s,e} = (\gamma_{s,e}(1), \dots, \gamma_{s,e}(T))'$, as

$$\gamma_{s,e}(t) = \begin{cases} \left\{ \frac{1}{12}(e-s+1)(e^2-2es+2e+s^2-2s) \right\}^{-1/2} \left(t - \frac{e+s}{2}\right), & t = s, \dots, e \\ 0, & \text{otherwise} \end{cases}. \quad (2.8)$$

Then we have that $\phi_{s,e}^b$ is orthonormal to both $\mathbf{1}_{s,e}$ and $\gamma_{s,e}$ (note that $\gamma_{s,e}$ itself is orthonormal to $\mathbf{1}_{s,e}$). The orthonormality of the vectors $\mathbf{1}_{s,e}$, $\gamma_{s,e}$ and $\phi_{s,e}^b$ is important in deriving the identity $\sigma_0^2 \mathcal{R}_{s,e}^b(\mathbf{Y}) = \mathcal{C}_{s,e}^b(\mathbf{Y})^2$ below, and helps improve the numerical efficiency and stability in our implementation of NOT. In particular, it means that the contrast function is invariant to both mean shifts and slope shifts on a given interval. In fact, $\phi_{s,e}^b$ can be derived by (i) applying the Gram–Schmidt process on the following vector (linear with a kink at $b+1$ on $[s, e]$)

$$\tilde{\phi}_{s,e}^b(t) = \begin{cases} t - b, & t = b+1, \dots, e \\ 0, & \text{otherwise} \end{cases}$$

with respect to $\mathbf{1}_{s,e}$ and $\gamma_{s,e}$, and (ii) normalisation such that $\|\cdot\|_2 = 1$.

Now write the restriction of \mathbf{v} on the interval $[s, e]$ as $\mathbf{v}|_{[s,e]} = (0, \dots, 0, v_s, \dots, v_e, 0, \dots, 0)'$. Fix any (s, e, b) , given the restriction imposed on Θ in (S2), the best approximation of $\mathbf{Y}|_{[s,e]}$ (in the ℓ_2 distance) with a single kink at b is a linear combination of $\mathbf{1}_{s,e}$, $\gamma_{s,e}$ and $\phi_{s,e}^b$ (all mutually orthonormal). Therefore,

$$\begin{aligned} \sigma_0^2 \mathcal{R}_{s,e}^b(\mathbf{Y}) &= \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{Y}|_{[s,e]} - a_0 \mathbf{1}_{s,e} - a_1 \gamma_{s,e}\|_2^2 - \min_{a_0, a_1, a_2 \in \mathbb{R}} \|\mathbf{Y}|_{[s,e]} - a_0 \mathbf{1}_{s,e} - a_1 \gamma_{s,e} - a_2 \phi_{s,e}^b\|_2^2 \\ &= \|\mathbf{Y}|_{[s,e]} - \langle \mathbf{Y}, \gamma_{s,e} \rangle \gamma_{s,e} - \langle \mathbf{Y}, \mathbf{1}_{s,e} \rangle \mathbf{1}_{s,e}\|_2^2 - \|\mathbf{Y}|_{[s,e]} - \langle \mathbf{Y}, \phi_{s,e}^b \rangle \phi_{s,e}^b - \langle \mathbf{Y}, \gamma_{s,e} \rangle \gamma_{s,e} - \langle \mathbf{Y}, \mathbf{1}_{s,e} \rangle \mathbf{1}_{s,e}\|_2^2 \\ &= \langle \mathbf{f}, \phi_{s,e}^b \rangle^2 = \mathcal{C}_{s,e}^b(\mathbf{Y})^2, \end{aligned}$$

i.e. the aforementioned requirements for the contrast function are satisfied.

2.3.3 Scenario (S3)

Here f_t is a piecewise-linear but not necessarily continuous function. We use the following contrast function for any $s < b < e$:

$$\mathcal{C}_{s,e}^b(\mathbf{v}) = \left(\langle \mathbf{v}, \psi_{s,e}^b \rangle^2 + \langle \mathbf{v}, \gamma_{s,b} \rangle^2 + \langle \mathbf{v}, \gamma_{b+1,e} \rangle^2 - \langle \mathbf{v}, \gamma_{s,e} \rangle^2 \right)^{1/2}. \quad (2.9)$$

This construction is justified by noting that

$$\begin{aligned} \sigma_0^2 \mathcal{R}_{s,e}^b(\mathbf{Y}) &= \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{Y}|_{[s,e]} - a_0 \mathbf{1}_{s,e} - a_1 \gamma_{s,e}\|_2^2 \\ &\quad - \left(\min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{Y}|_{[s,b]} - a_0 \mathbf{1}_{s,b} - a_1 \gamma_{s,b}\|_2^2 + \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{Y}|_{[b+1,e]} - a_0 \mathbf{1}_{b+1,e} - a_1 \gamma_{b+1,e}\|_2^2 \right) \\ &= \mathcal{C}_{s,e}^b(\mathbf{Y})^2, \end{aligned}$$

where we also used the orthonormality among $\mathbf{1}_{s,e}$, $\psi_{s,e}^b$, $\gamma_{s,b}$ and $\gamma_{b+1,e}$ in the above derivation.

2.3.4 Scenario (S4)

Here both f_t and σ_t are piecewise-constant. For any $1 \leq s+1 < b < e-1 \leq T$, we propose

$$\mathcal{C}_{s,e}^b(\mathbf{Y}) = (b-s+1) \log(\hat{\sigma}_{s,b}(\mathbf{Y})) + (e-b) \log(\hat{\sigma}_{b+1,e}(\mathbf{Y})) - (e-s+1) \log(\hat{\sigma}_{s,e}(\mathbf{Y})), \quad (2.10)$$

where

$$\hat{\sigma}_{s,e}^2(\mathbf{Y}) = \frac{1}{e-s+1} \sum_{t=s}^e \left(Y_t - \frac{1}{e-s+1} \sum_{t=s}^e Y_t \right)^2 = \langle \mathbf{Y}^2, \mathbf{1}_{s,e}^2 \rangle - \langle \mathbf{Y}, \mathbf{1}_{s,e}^2 \rangle^2.$$

Otherwise, for $b \notin \{s+2, \dots, e-2\}$, we set $\mathcal{C}_{s,e}^b(\mathbf{Y}) = 0$. In this Scenario, it is straightforward to verify that $\mathcal{C}_{s,e}^b(\mathbf{Y}) = \mathcal{R}_{s,e}^b(\mathbf{Y})$. (N.B. $\mathbf{1}_{s,e}^2 \neq \mathbf{1}_{s,e}$ because of the normalising constant.)

2.4 The NOT algorithm

Algorithm 1 NOT

Input: Data vector $\mathbf{Y} = (Y_1, \dots, Y_T)'$, F_T^M being a set of M intervals, with start- and end-points drawn independently and uniformly with replacement from $\{1, \dots, T\}$, $\mathcal{S} = \emptyset$.

Output: Set of estimated change-points $\mathcal{S} \subset \{1, \dots, T\}$.

procedure NOT(s, e, ζ_T)

if $e-s < 1$ **then** STOP

else

$\mathcal{M}_{s,e} := \{m : [s_m, e_m] \in F_T^M, [s_m, e_m] \subset [s, e]\}$

if $\mathcal{M}_{s,e} = \emptyset$ **then** STOP

else

$\mathcal{O}_{s,e} := \{m \in \mathcal{M}_{s,e} : \max_{s_m \leq b \leq e_m} \mathcal{C}_{s_m, e_m}^b(\mathbf{Y}) > \zeta_T\}$

if $\mathcal{O}_{s,e} = \emptyset$ **then** STOP

else

$m^* := \operatorname{argmin}_{m \in \mathcal{O}_{s,e}} |e_m - s_m|$

$b^* := \operatorname{argmax}_{s_{m^*} \leq b \leq e_{m^*}} \mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{Y})$

$\mathcal{S} := \mathcal{S} \cup \{b^*\}$

 NOT(s, b^*, ζ_T)

 NOT($b^* + 1, e, \zeta_T$)

end if

end if

end if

end procedure

Here we present a generic version of the NOT algorithm, which is described using pseudo-code in Algorithm 1 below. The main ingredient of the NOT procedure is a contrast function $\mathcal{C}_{s,e}^b(\cdot)$, chosen by the user, depending on the assumed nature of change-points in the data, e.g. as exemplified by our scenarios (S1)–(S4) above. The threshold $\zeta_T > 0$ is a tuning

parameter for the method with respect to which the contrast should be tested, while M is the number of the intervals drawn in the procedure. Guidance on the choice of ζ_T and M is given in Section 3.

2.5 Theoretical properties of NOT

In this section, we analyse the theoretical behaviour of the NOT algorithm in scenarios (S1) and (S2). An attractive feature of our methodology is that proofs for other scenarios can in principle be constructed “at home” by the user, by following the same generic proof strategy as the one we use for these two scenarios. A discussion of the proof strategy, as well its extensions, is given in Appendix B.2.

First, we revisit the canonical change-point detection problem, (S1), where the signal vector $\mathbf{f} = (f_1, \dots, f_T)'$ is piecewise-constant. For notational convenience, we set $\sigma_0 = 1$. Again σ_0 is assumed to be known. (If not, one can plug in the MAD estimator, described in Section 2.1, without affecting the correctness of our theory.)

Theorem 1. *Suppose Y_t follow model (2.1) in Scenario (S1). Let $\delta_T = \min_{j=1, \dots, q+1} (\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$, $\underline{f}_T = \min_{j=1, \dots, q} \Delta_j^{\mathbf{f}}$. Furthermore, assume that $\delta_T^{1/2} \underline{f}_T \geq \underline{C} \sqrt{\log T}$ for some large enough \underline{C} . Let \hat{q} and $\hat{\tau}_1, \dots, \hat{\tau}_q$ denote, respectively, the number and locations of change-points, sorted in increasing order, estimated by Algorithm 1 with the contrast function given by (2.4). Then there exist constants $C_1, C_2, C_3, C_4 > 0$ (all not depending on T) such that given $C_1 \sqrt{\log T} \leq \zeta_T < C_2 \delta_T^{1/2} \underline{f}_T$, $M \geq 36T^2 \delta_T^{-2} \log(T^2 \delta_T^{-1})$, and for sufficiently large T ,*

$$\mathbb{P} \left(\hat{q} = q, \max_{j=1, \dots, q} (|\hat{\tau}_j - \tau_j| (\Delta_j^{\mathbf{f}})^2) \leq C_3 \log T \right) \geq 1 - C_4/T. \quad (2.11)$$

In the simplest case where we have finitely many change-points with $\delta_T \sim T$, we need $M = O(\log T)$ many random intervals for the consistent detection of all the change-points, which leads to a total computational cost of $O(T \log T)$ for the entire procedure. Furthermore, $\max_{j=1, \dots, q} (|\hat{\tau}_j - \tau_j|) = O_P(\log T)$, which trails the minimax rate of $O_p(1)$ by only a logarithmic factor.

In addition, we note that the NOT procedure allows for $\delta_T^{1/2} \underline{f}_T$, a quantity that characterises the difficulty level of the problem, to be of order $\sqrt{\log T}$. As argued in Chan and Walther (2013) and Fryzlewicz (2014), this is the smallest rate that permits change-point detection for any method, and is thus optimal.

Next, we revisit Scenario (S2), in which the signal is piecewise-linear and continuous. Again, we set $\sigma_0 = 1$ for notational convenience.

Theorem 2. *Suppose Y_t follow model (2.1) in Scenario (S2). Let $\delta_T = \min_{j=1, \dots, q+1} (\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$, $\underline{f}_T = \min_{j=1, \dots, q} \Delta_j^{\mathbf{f}}$. Furthermore, assume that $\delta_T^{3/2} \underline{f}_T \geq \underline{C} \sqrt{\log T}$ for some large enough \underline{C} . Let \hat{q} and $\hat{\tau}_1, \dots, \hat{\tau}_q$ denote, respectively, the number and locations of change-points, sorted in increasing order estimated by Algorithm 1 with the contrast function given by (2.7). Then there exist constants $C_1, C_2, C_3, C_4 > 0$ not depending on T such that given $C_1 \sqrt{\log T} \leq \zeta_T < C_2 \delta_T^{3/2} \underline{f}_T$, $M \geq 36T^2 \delta_T^{-2} \log(T^2 \delta_T^{-1})$, and for sufficiently large T ,*

$$\mathbb{P} \left(\hat{q} = q, \max_{j=1, \dots, q} (|\hat{\tau}_j - \tau_j| (\Delta_j^{\mathbf{f}})^{2/3}) \leq C_3 (\log T)^{1/3} \right) \geq 1 - C_4/T. \quad (2.12)$$

In the case where we have finitely many change-points with $\delta_T \sim T$, we again need $M = O(\log T)$ many random intervals for the consistent estimation of all the change-points, leading to the total computational cost of $O(T \log T)$. In the most common case of $\underline{f}_T \sim T^{-1}$ (in which the signal f_t is bounded), the resulting change-point detection rate is $O_p(T^{2/3}(\log T)^{1/3})$, which is different from the minimax rate of $O_p(T^{2/3})$ derived by Raimondo (1998) by only a logarithmic factor. Moreover, in more general cases, the difficulty level of the problem in Scenario (S2) can be characterised by $\delta_T^{3/2} \underline{f}_T$, a quantity analogous to $\delta_T^{1/2} \underline{f}_T$ in the setting of (S1).

Finally, we remark that results similar to Theorem 1 and Theorem 2 can be obtained if we replace the assumption of standard Gaussian noise by $\mathbb{E}(\exp(u\varepsilon_t)) < \infty$ for some $u > 0$. In essence, we only require the tails of ε_t to be about or lighter than exponential, which can be seen from Step One and Step Two of the proofs in Appendix B.2 and Appendix B.3.

3 Computational aspects

3.1 Computing contrast functions in linear time

The practical performance (in terms of computational cost) of Algorithm 1 relies on the fast computation of the contrast functions discussed in Section 2.3 on any given interval $[s, e]$. In this section, we show that in all scenarios listed in Section 2.3, the cost of computing $\{\mathcal{C}_{s,e}^b(\mathbf{Y})\}_{b=s}^{e-1}$ is $O(e - s + 1)$.

Note that the key ingredients in $\mathcal{C}_{s,e}^b(\mathbf{Y})$ under the different scenarios are functions of the inner products, i.e. $\langle \mathbf{Y}, \phi_{s,e}^b \rangle$, $\langle \mathbf{Y}, \psi_{s,e}^b \rangle$, $\langle \mathbf{Y}, \gamma_{s,b} \rangle$, $\langle \mathbf{Y}, \gamma_{b+1,e} \rangle$, $\langle \mathbf{Y}, \mathbf{1}_{s,b}^2 \rangle$, $\langle \mathbf{Y}, \mathbf{1}_{b+1,e}^2 \rangle$, $\langle \mathbf{Y}^2, \mathbf{1}_{s,b}^2 \rangle$ and $\langle \mathbf{Y}^2, \mathbf{1}_{b+1,e}^2 \rangle$ for $b = s, \dots, e - 1$. For a fixed interval $[s, e]$, by simple algebra, we observe that $\langle \mathbf{Y}, \phi_{s,e}^b \rangle$ and $\langle \mathbf{Y}, \psi_{s,e}^b \rangle$ can be decomposed as

$$\begin{aligned} \langle \mathbf{Y}, \phi_{s,e}^b \rangle &= \overleftarrow{a}_{\phi,b} \sum_{t=s}^b Y_t - \overrightarrow{a}_{\phi,b} \sum_{t=b+1}^e Y_t \\ &:= \overleftarrow{a}_{\phi,b} \overleftarrow{\pi}_b^{(0)}(\mathbf{Y}) - \overrightarrow{a}_{\phi,b} \overrightarrow{\pi}_b^{(0)}(\mathbf{Y}), \\ \langle \mathbf{Y}, \psi_{s,e}^b \rangle &= \overleftarrow{a}_{\psi,b}^{(1)} \sum_{t=s}^b tY_t - \overrightarrow{a}_{\psi,b}^{(1)} \sum_{t=b+1}^e tY_t + \overleftarrow{a}_{\psi,b}^{(0)} \sum_{t=s}^b Y_t - \overrightarrow{a}_{\psi,b}^{(0)} \sum_{t=b+1}^e Y_t \\ &:= \overleftarrow{a}_{\psi,b}^{(1)} \overleftarrow{\pi}_b^{(1)}(\mathbf{Y}) - \overrightarrow{a}_{\psi,b}^{(1)} \overrightarrow{\pi}_b^{(1)}(\mathbf{Y}) + \overleftarrow{a}_{\psi,b}^{(0)} \overleftarrow{\pi}_b^{(0)}(\mathbf{Y}) - \overrightarrow{a}_{\psi,b}^{(0)} \overrightarrow{\pi}_b^{(0)}(\mathbf{Y}), \end{aligned}$$

where $\overleftarrow{a}_{\phi,b}$, $\overrightarrow{a}_{\phi,b}$, $\overleftarrow{a}_{\psi,b}^{(1)}$, $\overrightarrow{a}_{\psi,b}^{(1)}$, $\overleftarrow{a}_{\psi,b}^{(0)}$ and $\overrightarrow{a}_{\psi,b}^{(0)}$ are scalars that do not depend on \mathbf{Y} , and can all be computed at the cost of $O(1)$ using equations given in Section 2.3. Here for notational convenience, we use overhead arrows to indicate whether a scalar or a function is associated with observations to the left of b (i.e. $[s, b]$, using $\overleftarrow{\cdot}$) or to the right of b (i.e. $[b+1, e]$, using $\overrightarrow{\cdot}$). We also suppress their dependence on s and e in the notation. In addition, the following recursive formulae hold

$$\begin{aligned} \overleftarrow{\pi}_{b+1}^{(k)}(\mathbf{Y}) &= \overleftarrow{\pi}_b^{(k)}(\mathbf{Y}) + (b+1)^k Y_{b+1}, \\ \overrightarrow{\pi}_b^{(k)}(\mathbf{Y}) &= \overrightarrow{\pi}_{b+1}^{(k)}(\mathbf{Y}) + (b+1)^k Y_{b+1}, \end{aligned}$$

with $\overleftarrow{\pi}_s^{(k)}(\mathbf{Y}) = \overrightarrow{\pi}_e^{(k)}(\mathbf{Y}) = 0$ for $k = 0, 1$. Consequently, $\overleftarrow{\pi}_b^{(k)}(\mathbf{Y})$ and $\overrightarrow{\pi}_b^{(k)}(\mathbf{Y})$ for all $b \in \{s, \dots, e-1\}$ and $k = 0, 1$ (thereby $\langle \mathbf{Y}, \phi_{s,e}^b \rangle$ and $\langle \mathbf{Y}, \psi_{s,e}^b \rangle$) can be computed in a single pass through Y_s, \dots, Y_e . Similar approach can be applied to the remaining inner products involved in the definitions of the contrast functions given in Section 2.3, which demonstrates that in all these cases the computation of $\{\mathcal{C}_{s,e}^b(\mathbf{Y})\}_{b=s}^{e-1}$ scales linearly with the number of observations.

3.2 The NOT solution path algorithm

In general, there are at least two ways of choosing a suitable threshold ζ_T in Algorithm 1. It can be either done by selecting a ζ_T which guarantees consistent change-point estimation in a given class of segmentation problems with a high probability, or by using one that optimises a loss function or a model selection criterion. The latter approach proves particularly useful when the theoretically “optimal” threshold is either difficult to derive, or depends on some unobserved quantities, which is typically the case. Denote by $\mathcal{T}(\zeta_T) = \{\hat{\tau}_1(\zeta_T), \dots, \hat{\tau}_{\hat{q}}(\zeta_T)\}$ the locations of change-points estimated by Algorithm 1 with threshold ζ_T (where we suppress the dependence of \hat{q} on ζ_T for notational convenience) and define the solution path as the family of sets $\{\mathcal{T}(\zeta_T)\}_{\zeta_T \geq 0}$. In this section, we present a fast algorithm that computes the entire solution path of Algorithm 1. Being able to compute the solution path quickly is essential in Section 3.4, where we study a data-driven approach to the choice of ζ_T .

The solution path seen as the function $\zeta_T \mapsto \mathcal{T}(\zeta_T)$ changes only at discrete points, i.e. there exist $0 \leq \zeta_T^{(1)} < \dots < \zeta_T^{(N)}$, such that $\mathcal{T}(\zeta_T^{(i)}) \neq \mathcal{T}(\zeta_T^{(i+1)})$ for any $i = 1, \dots, N-1$, and $\mathcal{T}(\zeta_T) = \mathcal{T}(\zeta_T^{(i)})$ for any $\zeta_T \in [\zeta_T^{(i)}, \zeta_T^{(i+1)})$. Furthermore, we have that $\mathcal{T}(\zeta_T) = \emptyset$ for any $\zeta_T \geq \zeta_T^{(N)}$. Thresholds $\zeta_T^{(i)}$ are unknown and depend on the data, therefore applying Algorithm 1 on a range of pre-specified thresholds typically does not recover the entire solution path. From the computational point of view, repeated application of Algorithm 1 to find the solution path is not optimal either, because intuitively one would expect the solutions for $\zeta_T^{(i+1)}$ and $\zeta_T^{(i)}$ to be similar for most i .

Below we propose our Algorithm 2 that computes the entire solution path $\{\mathcal{T}(\zeta_T)\}_{\zeta_T \geq 0}$. Its construction stems from the following two observations. First, for any fixed threshold ζ_T , Algorithm 1 implies a binary tree data structure that is constructed according to the order of the detection of each change-point. More specifically, in our implementation, each tree node \mathbf{N} contains information on the location of the detected change-point $\mathbf{N.b}$ over the interval of interest, $[\mathbf{N.s}, \mathbf{N.e}]$, along with the maximum achieved value of the contrast function over all intervals in F_T^M that are subsets of $[\mathbf{N.s}, \mathbf{N.e}]$ (the largest value and its location are denoted by $\mathbf{N.c}$ and $\mathbf{N.b}$, respectively). Moreover, we define $\mathbf{N.Left}$ and $\mathbf{N.Right}$ pointing to the nodes of the next detected change-points in $[\mathbf{N.s}, \mathbf{N.b}]$ and $[\mathbf{N.b} + 1, \mathbf{N.e}]$, respectively. We then treat the first detected change-point over $[1, T]$ as the root of the tree and construct its branches in a recursive fashion afterwards. Second, suppose that we have already constructed the tree for ζ_T with root $\mathbf{N_r}$. For $\zeta'_T > \zeta_T$, the new tree’s root is unchanged if $\mathbf{N_r.c} > \zeta'_T$. This observation remains valid for $\mathbf{N_r.Left}$ and $\mathbf{N_r.Right}$ and all subsequent nodes. Therefore, a branch of the tree has to be reconstructed only if $\mathbf{N.c} \leq \zeta'_T$ for some node \mathbf{N} . In this way, the tree constructed for ζ_T can be used as a starting point to finding the tree corresponding to ζ'_T , thus significantly reducing the computational time in comparison to constructing the

Algorithm 2 NOT solution path

Input: Intervals $[s_m, e_m]$ and

$$b_m := \operatorname{argmax}_{s_m \leq b \leq e_m} \mathcal{C}_{s_m, e_m}^b(\mathbf{Y}), \quad c_m := \mathcal{C}_{s_m, e_m}^{b_m}(\mathbf{Y}), \quad l_m := e_m - s_m + 1$$

for all $m \in F_T^M$.

Output: Thresholds $0 = \zeta_T^{(1)} < \dots < \zeta_T^{(N)}$ and sets of estimated change-points $\mathcal{T}(\zeta_T^{(1)}), \dots, \mathcal{T}(\zeta_T^{(N)})$.

```
procedure BUILDBINARYTREE( $s, e, \zeta_T, \mathbf{N}$ )  
   $\mathcal{M}_{s,e} :=$  set of those  $m \in \{1, \dots, M\}$  such that  $[s_m, e_m] \subset [s, e]$   
   $\mathcal{O}_{s,e} :=$  set of  $m \in \mathcal{M}_{s,e}$  such that  $c_m > \zeta_T$   
  if  $\mathcal{O}_{s,e} = \emptyset$  then  $\mathbf{N} = \text{NULL}$   
  else  
     $k :=$  any elements of  $\operatorname{argmin}_{m \in \mathcal{O}_{s,e}} l_m$   
     $\mathbf{N.b} := b_k, \mathbf{N.c} := c_k, \mathbf{N.Left} := \text{NULL}, \mathbf{N.Right} := \text{NULL}$   
    BUILDBINARYTREE( $s, \mathbf{N.b}, \zeta_T, \mathbf{N.Left}$ )  
    BUILDBINARYTREE( $\mathbf{N.b} + 1, e, \zeta_T, \mathbf{N.Right}$ )  
  end if  
end procedure
```

```
procedure UPDATEBINARYTREE( $s, e, \zeta_T, \mathbf{N}$ )  
  if  $\mathbf{N.c} \leq \zeta_T$  then  
    BUILDBINARYTREE( $s, e, \zeta_T, \mathbf{N}$ )  
  else  
    if  $\mathbf{N.Left} \neq \text{NULL}$  then  
      UPDATEBINARYTREE( $s, \mathbf{N.b}, \zeta_T, \mathbf{N.Left}$ )  
    end if  
    if  $\mathbf{N.Right} \neq \text{NULL}$  then  
      UPDATEBINARYTREE( $\mathbf{N.b} + 1, e, \zeta_T, \mathbf{N.Right}$ )  
    end if  
  end if  
end procedure
```

```
procedure SOLUTIONPATH()  
  Set  $\mathbf{N_r} := \text{NULL}, i := 1, \zeta_T^{(1)} := 0$   
  BUILDBINARYTREE( $1, T, \zeta_T^{(1)}, \mathbf{N_r}$ )  
  while  $\mathbf{N_r} \neq \text{NULL}$  do  
     $\mathcal{D} := \{\mathbf{N_r} \text{ and all its children nodes}\}$   
     $\mathcal{T}(\zeta_T^{(i)}) := \{\mathbf{N.b} | \mathbf{N} \in \mathcal{D}\}$   
     $\zeta_T^{(i+1)} := \min_{\mathbf{N} \in \mathcal{D}} \{\mathbf{N.c}\}$   
    UPDATEBINARYTREE( $1, T, \zeta_T^{(i+1)}, \mathbf{R}$ )  
     $i := i + 1$   
  end while  
end procedure
```

tree from scratch. See the pseudo-code of Algorithm 2 for more details.

3.3 An illustrative example

In this part, we revisit the example shown in the Introduction, and provide a simple illustration of how Algorithm 1 and Algorithm 2 work on a simulated dataset. Figure 3 shows the generated data $\{Y_t\}_{t=1}^{1000}$ following Scenario (S2), where the signal f_t is as in (1.2) and $\sigma_t = 0.05$. The contrast function (2.7) is evaluated for 5 intervals. We observe that the contrast function corresponding to $[1, 1000]$, being the longest interval here, attains its maximum at $b = 490$, which is far from the true change-points located at $\tau = 350$ and $\tau = 650$. Furthermore, $\max_{1 \leq b \leq 1000} \mathcal{C}_{s,e}^b(\mathbf{Y})$ is much larger than the corresponding value for the other intervals considered in Table 1. However, thanks to the fact that we focus on the narrowest-over-threshold intervals, Algorithm 1 (for any $\zeta_T \in (0.08, 0.83)$) picks at its first iteration an interval with exactly one change-point (depending on ζ_T , it is either $[225, 450]$ or $[500, 750]$) and the maximum of the contrast function computed is close to one of the true change-points.

s	e	$e - s + 1$	$\operatorname{argmax}_{s \leq b \leq e} \mathcal{C}_{s,e}^b(\mathbf{Y})$	$\max_{s \leq b \leq e} \mathcal{C}_{s,e}^b(\mathbf{Y})$
1	1000	1000	490	10.19
10	245	236	43	0.08
225	450	226	344	0.76
500	750	251	651	0.83
740	950	211	746	0.03
450	550	101	471	0.07

Table 1: Intervals considered in Figure 3(a) and corresponding maxima of the contrast function $\mathcal{C}_{s,e}^b(\mathbf{Y})$ given by (2.7), all calculated for a sample path of Y_t , $t = 1, \dots, 1000$ generated from model (1.1) with the signal f_t given by (1.2) and the noise $\varepsilon_t \sim \mathcal{N}(0, 0.05^2)$.

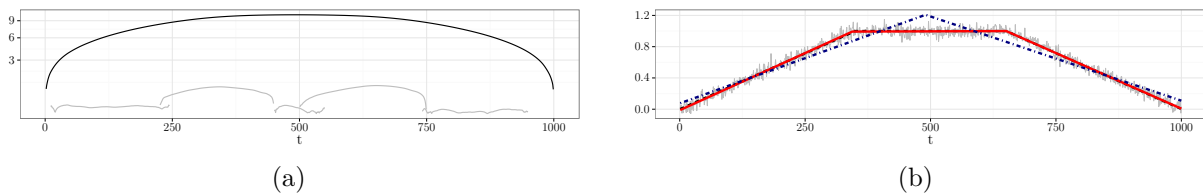


Figure 3: An application of the NOT methodology to Y_t generated from model (1.1) with the signal f_t given by (1.2) and i.i.d. $\varepsilon_t \sim \mathcal{N}(0, 0.05^2)$. Figure 3(a): contrast function $\mathcal{C}_{s,e}^b(\mathbf{Y})$ given by (2.7) evaluated for all $b \in [s, e]$ and intervals $[s, e]$ specified in Table 1. For intervals containing one change-point, $\mathcal{C}_{s,e}^b(\mathbf{Y})$ attains its maximum at b close to the change-point. When there are two change-points (black solid line), the maximum is far from both change-points, despite $\max_{s \leq b \leq e} \mathcal{C}_{s,e}^b(\mathbf{Y})$ being large. Figure 3(b): observed Y_t (thin grey), true signal (thick dashed black), signal estimated picking the change-point candidate based on the interval corresponding to the largest contrast function (dotted-dashed navy) and the *narrowest-over-threshold* intervals (dashed red).

Figure 4 shows how Algorithm 2 proceeds in the example presented in Figure 3. At the initial stage that can be seen in Figure 4(a), the threshold is set to $\zeta_T^{(1)} = 0$ and $b = 417$, the maximum of the contrast function computed for the shortest interval $[450, 550]$ is taken as the root of the binary tree. Then we construct its left and right branches by considering only those intervals specified in Table 1 whose endpoints $[s, e] \subset [1, 471]$ and $[s, e] \subset [472, 1000]$, respectively, and the procedure continues for the resulting nodes. Next, the node with the smallest value of the contrast function is determined ($b = 746$) and the threshold is set to the corresponding minimum $\zeta_T^{(2)} = 0.03$. This guarantees that as Algorithm 2 proceeds, there will be at least one update in the binary tree. In our example, the $b = 746$ node is removed and, as the maximum for $[500, 750] \subset [472, 1000]$ exceeds the threshold, the $b = 651$ node is inserted its place. Subsequently, we identify the node with the smallest contrast again ($b = 471$), update the threshold to $\zeta_T^{(3)} = 0.07$ and reconstruct the entire tree, as $b = 471$ in Figure 4(b) constitutes its root. Algorithm 2 keeps running until the resulting tree shrinks to NULL. In this example, the fourth solution on the path (Figure 4(d)) contains exactly two nodes being close to the true change-points.

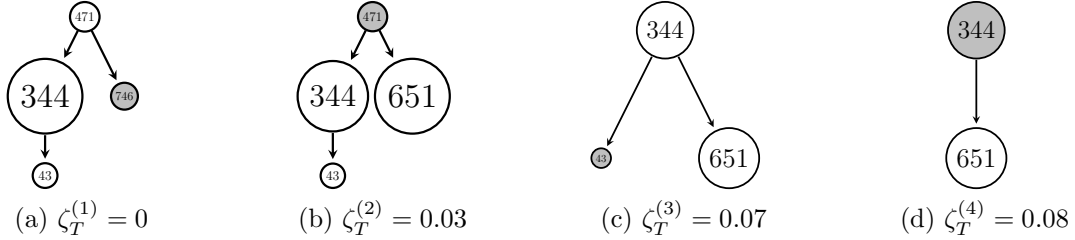


Figure 4: First four segmentation trees obtained by Algorithm 2 applied to a Y_1, \dots, Y_{1000} presented in Figure 3. The larger the node, the larger the corresponding value of $\max_{s \leq b \leq e} \mathcal{C}_{s,e}^b(\mathbf{Y})$ given by (2.7). The grey nodes correspond to the smallest contrast function for each tree and are updated as Algorithm 2 proceeds.

3.4 Parameter choice

3.4.1 Choice of M

We recommend setting $M = 10000$ when the number of observations is of the order of thousands. Our empirical evidence shows that setting a much higher M does not improve the practical performance of the method in these circumstances. With this value of M , the implementation of Algorithm 1 provided in the **R not** package (Baranowski et al., 2016b) achieves the average computation time not longer than 2 seconds in all examples discussed in Section A using a single core of an Intel Xeon 3.6 GHz CPU. This can be accelerated further, as the **not** package allows for computing the contrast function over the intervals drawn in parallel using all available CPU cores.

3.4.2 Choice of the threshold ζ_T

Algorithm 1 can be applied to a wide range of change-point detection problems with various contrast functions, hence it seems challenging (at least from a theoretical perspective) to find

a universal threshold that works well in all settings. In the piecewise-constant and piecewise-linear cases, based on Theorem 1 and Theorem 2, respectively, we could take ζ_T of the lowest admissible order (i.e. $\sqrt{\log T}$). Here, our ambition is to come up with a more general data-driven choice of ζ_T based on Algorithm 2. Let $\mathcal{T}(\zeta^{(1)}), \dots, \mathcal{T}(\zeta^{(N)})$ be the NOT solution path, i.e. the collection of candidate models produced by Algorithm 2. We propose to select $\mathcal{T}(\zeta^{(k)})$ minimising the Schwarz Information Criterion (SIC) defined as follows. Let $k = 1, \dots, N$, $\hat{q}_k = |\mathcal{T}(\zeta_T^{(k)})|$ and $\hat{\Theta}_1, \dots, \hat{\Theta}_{\hat{q}_k+1}$ be the maximum likelihood estimators of the segment parameters in model (2.1) with the estimated change-points $\hat{\tau}_1, \dots, \hat{\tau}_{\hat{q}_k} \in \mathcal{T}(\zeta_T^{(k)})$. Denote by n_k the total number of estimated parameters, including the number of free parameters in $\Theta_1, \dots, \Theta_{\hat{q}_k+1}$ (this can be different from the dimensionality of each Θ_j multiplied by the number of segments, as e.g. in (S2)), and \hat{q}_k . The SIC criterion is given by

$$\text{SIC}(k) = -2 \sum_{j=1}^{\hat{q}_k+1} \ell(Y_{\hat{\tau}_{j-1}+1}, \dots, Y_{\hat{\tau}_j}; \hat{\Theta}_j) + n_k \log(T), \quad (3.1)$$

with $\hat{\tau}_0 = 0$ and $\hat{\tau}_{\hat{q}_k+1} = T$. In practice it may not be necessary to calculate SIC for all k , if the number of change-points in the data is expected to be rather moderate. In all applications presented in this work we compute SIC only for k such that $|\mathcal{T}(\zeta_T^{(k)})| \leq q_{\max}$ with $q_{\max} = 25$. In general, solutions on the path corresponding to very small values of ζ_T contain many estimated change-points, especially when M is large. Such solutions are unlikely to minimise (3.1), therefore by considering $|\mathcal{T}(\zeta_T^{(k)})| \leq q_{\max}$ we achieve computational gains, without adversely impacting the overall performance of the methodology.

3.5 Computational complexity of the NOT and NOT solution path algorithms

Here we elaborate on the computational complexity of Algorithms 1 and 2. For both algorithms, the task of computation can be divided into two main parts. First, we need to evaluate a chosen contrast function for all points in the M randomly picked intervals with their endpoints in $\{1, \dots, T\}$. In the second part, we find potential locations of the change-points for a single threshold ζ_T in the case of Algorithm 1 and for all possible thresholds in the case of Algorithm 2.

Naturally, the total computational complexity of the first part depends on the cost of computing the contrast function for a single interval. In all scenarios studied in this paper, this cost is linear in the length of an interval, as shown in Section 3.1. The intervals drawn in the procedures have approximately $T/4$ points on average, therefore the computational complexity of the first part of the computations is $O(MT)$ in a typical application. Importantly, as the calculations for one interval are completely independent of the calculations for another, it is straightforward to run these computations in parallel. Implementation of the NOT methodology available from the R package **not** (Baranowski et al., 2016b) uses to this end the **OpenMP** framework (Dagum and Menon, 1998), allowing for the efficient use of multiple cores that modern CPUs offer.

As we explain in Section 3.2, finding solutions of Algorithm 1 for a single threshold ζ_T is equivalent to the construction of a binary tree, which can be performed with the

BUILDBINARYTREE routine given in Algorithm 2. Computational cost of this operation is no larger than $O(MK_{\zeta_T})$, where K_{ζ_T} denotes the height of the constructed binary tree with the threshold ζ_T . The computational complexity of finding the entire solution path using Algorithm 2 is therefore (in the worst case) of the order $O(MKN)$, where N and K are, respectively, the number of solutions and the maximum tree depth over the entire solution path. However, this is a rough estimate which assumes that for each threshold on the path the binary tree has a different root node, which, from our empirical experience, is highly unlikely to occur in practice. Typically, the consecutive trees on the path differ just slightly, see e.g. Figure 6, which significantly reduces the amount of computation that Algorithm 2 requires. Finally, we remark that the memory complexity of Algorithm 2 is $O(MT)$, which combined with its low computational complexity implies that our approach can handle problem of size T in the millions.

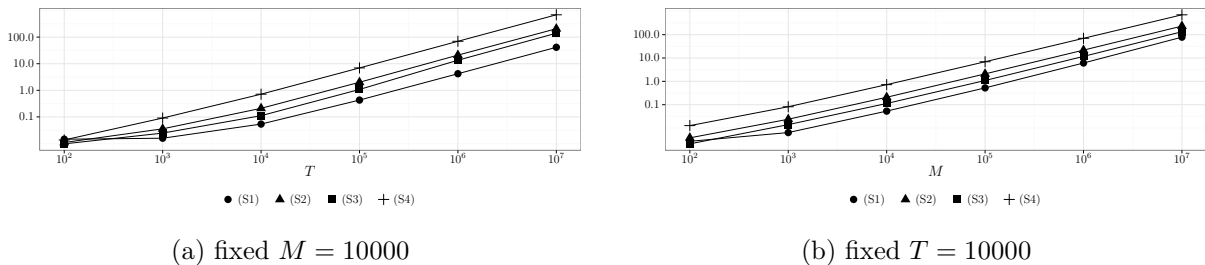


Figure 5: Execution times (in seconds) for the implementation of Algorithm 2 available from R package **not** (Baranowski et al., 2016b), for various feature detection problems with the data Y_t , $t = 1, \dots, T$ being i.i.d. $\mathcal{N}(0, 1)$. In a single run, computations for the input of the algorithm are performed in parallel, using 8 virtual cores of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM. The computation times are averaged over 10 runs in each case.

Figure 5 shows execution times for the implementation of Algorithm 2 available from the R package **not**, with the data Y_t , $t = 1, \dots, T$, being i.i.d. $\mathcal{N}(0, 1)$. The running times appears to scale linearly both in T (Figure 5(b)) and in M (Figure 5(b)), which provides evidence that the computational complexity of Algorithm 2 in this example is practically of the order $O(MT)$.

4 Simulation study

We compare the performance of the R package **not** implementing the NOT methodology against the best competitors available on CRAN. The R code for all simulations can be downloaded from our GitHub repository (Baranowski et al., 2016a). We consider examples following (S1)–(S4) introduced in Section 2.3, as well as an extra example satisfying

(S5) $\sigma_t = \sigma_0$ and f_t is a piecewise-quadratic function of t .

Calculations required to derive the contrast function in (S5) are similar to those shown in Section 2.3 for (S3); we omit them here.

To the best of our knowledge, none of the competing packages can be applied in all of the scenarios (S1)–(S5). For change-point detection in the mean, the competitors are: **changepoint** (Killick and Eckley, 2014) implementing the PELT methodology proposed by Killick et al. (2012a), **changepoint.np** (Haynes et al., 2016b) implementing a nonparametric extension of the PELT methodology studied in Haynes et al. (2016a), **wbs** (Baranowski and Fryzlewicz, 2015) implementing the Wild Binary Segmentation proposed by Fryzlewicz (2014), **ecp** (James and Matteson, 2014) implementing the e.cp3o method proposed by James and Matteson (2015), **strucchange** (Zeileis et al., 2002) implementing the methodology of Bai and Perron (2003), **Segmentor3IsBack** (Cleynen et al., 2013) implementing the technique proposed by Rigaiil (2010), **nmcd** (Zou and Lancezhang, 2014), implementing the NMCD methodology of Zou et al. (2014), **stepR** (Hotz and Sieling, 2016), implementing the SMUCE method proposed by Frick et al. (2014). We refer to the corresponding methods as, respectively, PELT, NP-PELT, WBS, e.cp3o, B&P, S3IB, NMCD and SMUCE. All techniques but B&P, WBS, S3IB and SMUCE can be also used for change-point detection in (S4), where change-points occur in the mean and variance of the data.

Only the B&P method allows for change-point detection in piecewise-linear and piecewise-quadratic signals, hence we also study the performance of the trend filtering methodology of Kim et al. (2009) termed as TF hereafter, using the implementation available from the R package **genlasso** (Taylor and Tibshirani, 2014), to have a broader comparison. The TF method aims to estimate a piecewise polynomial signal from the data, not focusing on the change-point detection problem directly. Let $\hat{f}_t^{(TF)}$ denote the TF estimate of the true signal f_t , then the TF estimates of the change-points in (S2) are defined as those τ for which $|2\hat{f}_\tau^{(TF)} - \hat{f}_{\tau-1}^{(TF)} - \hat{f}_{\tau+1}^{(TF)}| > \epsilon$, where $\epsilon > 0$ is a very small number being the numerical tolerance level (more precisely, we set $\epsilon = 1.11 \times 10^{-15}$). In the piecewise-polynomial case, the change-points are defined as those τ for which the third order differences $|\hat{f}_{\tau+2}^{(TF)} - 3\hat{f}_{\tau+1}^{(TF)} + 3\hat{f}_\tau^{(TF)} - \hat{f}_{\tau-1}^{(TF)}| > \epsilon$. Finally, we note that both B&P and TF require a substantial amount of computational resources, with B&P being the slowest among all methods considered in this study. Owing to this, below we consider signals of moderate lengths not exceeding a few thousand, however, as demonstrated in Section 3.5, our proposal can be applied even if T is of the order of 10^7 .

In this section, we apply Algorithm 2 to compute the NOT solution path and always pick the solution minimising the SIC criterion introduced in Section 3.4. The number of intervals drawn in the procedure and the maximum number of change-points for SIC are set to $M = 10000$ and $q_{max} = 25$, respectively. In each simulated example, we use the contrast function designed to detect change-points in the scenario that the example follows, derived in Section 2.3 under the assumption that ε_t is Gaussian. The resulting method is referred to simply as ‘NOT’. The tuning parameters for the competing methods are set to the values recommended by the authors of the corresponding R packages.

The simulation results below show that the NOT methodology with the Gaussian contrast functions is fairly robust against the misspecification of the distribution of the noise. Nevertheless, to illustrate how its performance can be improved further in the presence of heavy-tailed noise, in simulation models for Scenario (S1) we apply Algorithm 2 with an additional contrast function, defined for \mathbf{Y} and $1 \leq s \leq b < e < T$ as

$$\mathcal{C}_{s,e}^b(\mathbf{Y}) = \langle \mathcal{S}_{s,e}(\mathbf{Y}), \psi_{s,e}^b \rangle, \quad (4.1)$$

where for any vector $\mathbf{v} = \text{asise}(v_1, \dots, v_T)'$ the i -component of $\mathcal{S}_{s,e}(\mathbf{v})$ is given by $\mathcal{S}_{s,e}(\mathbf{v})_i = \text{sign}(v_i - (e - s + 1)^{-1} \sum_{t=s}^e v_t)$ and $\boldsymbol{\psi}_{s,e}^b$ is defined by (2.3). The rationale behind (4.1) is as follows. Suppose Y_t satisfies (1.1) with the piecewise-constant signal f_t and let $[s, e]$ be any interval containing exactly one change-point at $\tau \in [s, e]$. For $i = s, \dots, e$, consider $\tilde{Y}_i = \text{sign}(Y_t - (e - s + 1)^{-1} \sum_{t=s}^e f_t)$. Then \tilde{Y}_i decomposes as $\tilde{Y}_i = \tilde{f}_t + \tilde{\varepsilon}_t$, where $\tilde{f}_t = \mathbb{E} \text{sign}(Y_t - (e - s + 1)^{-1} \sum_{t=s}^e f_t)$ also has exactly one change-point at τ , while the distribution of $\tilde{\varepsilon}_t$ is binomial (regardless of the distribution for the original noise ε_t), hence its tails are light. In this setting, as argued in Section 2.5, (2.4) can be used to identify the location of the change-point in $\tilde{Y}_s, \dots, \tilde{Y}_e$. As the true signal is unknown, we use $\bar{\mathbf{Y}}_{s,e} := (e - s + 1)^{-1} \sum_{t=s}^e Y_t$ as a proxy for $(e - s + 1)^{-1} \sum_{t=s}^e f_t$ when computing (2.4) for the data \mathbf{Y} . In essence, we assign $Y_s - \bar{Y}_{s,e}, \dots, Y_e - \bar{Y}_{s,e}$ (i.e. residuals for fitting a curve with no change-point on a given interval) into two classes (± 1), and apply the contrast function to their labels. Algorithm 2 combined with (4.1) and SIC is termed ‘NOT HT’, where ‘HT’ stands for heavy tails. We expect that the theoretical properties of NOT HT can be shown along the lines of Theorem 1, because the tails of $\tilde{\varepsilon}_t$ are lighter than exponential. Finally, we note that the contrast functions addressing the issue of heavy-tails in the noise can be also constructed for (S2)–(S5). For example, when the distribution of the noise is known, this can be achieved by considering GLR given by (2.2) with the correct likelihood function. Otherwise, on any given interval $[s, e]$, one could again consider the vector of residuals from fitting a corresponding curve with no change-point, and truncate the residuals on that interval by a small proportion before plugging it (instead of \mathbf{Y}) into the contrast function. This approach is robust, and intuitively preserves more information than using just the sign operator and could be useful for determining the location of a change-point in segments of a more complicated parametric form.

We simulate data according to equation (2.1) using the test signals (M1) **teeth**, (M2) **blocks**, (M3) **wave1**, (M4) **wave2**, (M5) **mix**, (M6) **vol** and (M7) **quad**, with the i.i.d. noise following $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 2)$, scaled Laplace or scaled Student- t_5 distributions. A detailed specification of which can be found in Appendix A. Figure 6 shows the examples of the data generated from models (M1)–(M7), as well as estimates produced by NOT.

Tables 2–5 summarise the results for the four different distributions of the noise ε_t . For each method, we show a frequency table for the distribution of $\hat{q} - q$, where \hat{q} is the number of the estimated change-points and q denotes the true number of change-points. We also report Monte-Carlo estimates of the Mean-Square error

$$\text{MSE} = \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left(f_t - \hat{f}_t \right)^2. \quad (4.2)$$

For all methods but TF, \hat{f}_t is calculated by finding the OLS approximation of the signal of the appropriate type depending on the true f_t , between each consecutive pair of estimated change-points. For TF, \hat{f}_t used in the definition of the MSE is the penalised least squares estimate of f_t returned by the TF algorithm. To assess the performance of each method in terms of the accuracy of the estimated locations of the change-points, we also report estimates of the (scaled) Hausdorff distance defined as

$$d_H = T^{-1} \mathbb{E} \max \left\{ \max_{j=0, \dots, \hat{q}+1} \min_{k=0, \dots, \hat{q}+1} |\tau_j - \hat{\tau}_k|, \max_{k=0, \dots, \hat{q}+1} \min_{j=0, \dots, \hat{q}+1} |\hat{\tau}_k - \tau_j| \right\}, \quad (4.3)$$

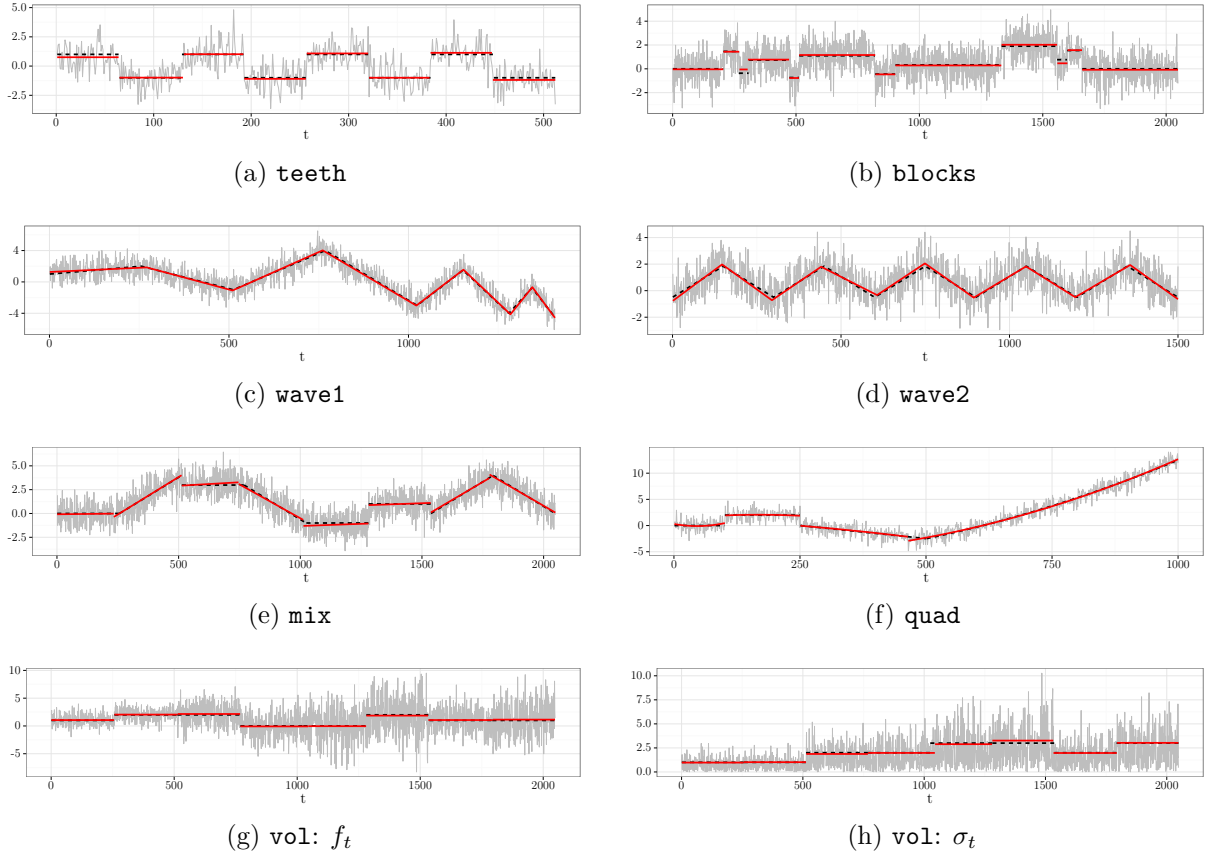


Figure 6: Examples of data generated from simulation models studied in Section A. Figure 6(a)– 6(g): data series Y_t (thin grey), true signal f_t (dashed black), \hat{f}_t being the OLS estimate of f_t with the change-points estimated by NOT (thick red). Figure 6(h): centered data $|Y_t - \hat{f}_t|$ (thick grey), true standard deviation σ_t (dashed black) and the estimated standard deviation $\hat{\sigma}_t$ between the change-points detected by NOT (thick red).

where $0 = \tau_0 < \tau_1 < \dots \tau_q < \tau_{q+1} = T$ and $0 = \hat{\tau}_0 < \hat{\tau}_1 < \dots \hat{\tau}_q < \hat{\tau}_{q+1} = T$ denote, respectively, true and estimated locations of the change-points. From the definition above, it follows that $0 \leq d_H \leq 1$. An estimator is regarded to perform well when its d_H is close to 0. However, when the number of change-points is under-estimated or some of the estimated change-points are not close to the real ones, d_H is closer to 1.

The points below, grouped according to the scenario for the type of segmentation problem, discuss the results.

- (S1) Two simulation models follow this scenario: (M1) **teeth** and (M2) **blocks**. The **teeth** signal with the $\mathcal{N}(0, 1)$ noise is a relatively easy setting, where all methods but B&P always detect all change-points. PELT, SMUCE and e-cp3o perform exceptionally well here, always finding exactly 7 change-points close to the true locations. NMCD, NOT, NOT HT, S3IB and WBS overestimate q sporadically, while NP-PELT shows a tendency of detecting some additional change-points. The performance of NP-PELT and SMUCE deteriorates in (M1) when $\varepsilon_t \sim \mathcal{N}(0, 2)$; SMUCE underestimates q , while

Method	Model	$\hat{q} - q$							MSE	$d_H \times 10^2$	time
		≤ -3	-2	-1	0	1	2	≥ 3			
B&P	(M1)	70	8	1	21	0	0	0	0.703	11.39	0.27
e-cp3o		0	0	0	100	0	0	0	0.052	0.48	2.32
NMCD		0	0	0	96	4	0	0	0.093	0.76	1.38
NOT		0	0	0	99	1	0	0	0.053	0.54	0.08
NOT HT		0	0	0	99	1	0	0	0.055	0.51	0.1
NP-PELT		0	0	0	86	11	2	1	0.068	0.85	0.03
PELT		0	0	0	100	0	0	0	0.052	0.48	0
S3IB		0	0	0	92	6	2	0	0.055	0.67	0.11
SMUCE		0	0	0	100	0	0	0	0.083	0.57	0.22
WBS		0	0	0	97	3	0	0	0.054	0.58	0.11
B&P	(M2)	100	0	0	0	0	0	0	0.314	12.56	4.29
e-cp3o		100	0	0	0	0	0	0	0.127	5.69	188.84
NMCD		0	5	64	31	0	0	0	0.035	1.82	4.92
NOT		0	4	61	35	0	0	0	0.026	1.56	0.11
NOT HT		2	8	54	28	8	0	0	0.033	2.08	0.23
NP-PELT		0	0	27	44	15	9	5	0.029	2.13	0.49
PELT		11	33	45	11	0	0	0	0.035	2.97	0.01
S3IB		0	2	49	49	0	0	0	0.024	1.42	0.51
SMUCE		59	36	5	0	0	0	0	0.069	3.44	0.03
WBS		0	1	45	53	0	1	0	0.026	1.31	0.22
B&P	(M3)	0	0	100	0	0	0	0	0.218	3.78	147.23
NOT		0	0	0	99	1	0	0	0.015	0.99	0.63
TF		0	0	0	0	0	0	100	0.019	8.33	63.98
B&P	(M4)	0	1	3	96	0	0	0	0.072	2.59	168.12
NOT		0	0	0	100	0	0	0	0.016	1.21	0.53
TF		0	0	0	0	0	0	100	0.016	4.3	64.81
B&P	(M5)	0	0	0	100	0	0	0	0.02	2.42	382.96
NOT		0	0	0	99	1	0	0	0.02	2.42	0.51
TF		0	0	0	0	0	0	100	0.026	6.03	77.09
e-cp3o	(M6)	94	3	0	3	0	0	0	0.378	16.83	11.35
NMCD		0	0	7	83	8	2	0	0.057	2.54	4.8
NOT		0	0	4	94	2	0	0	0.049	1.69	1.22
NP-PELT		0	0	0	20	30	19	31	0.123	2.96	0.61
PELT		9	15	28	48	0	0	0	0.074	8	0.02
B&P	(M7)	0	0	0	100	0	0	0	0.021	1.94	44.14
NOT		0	0	0	100	0	0	0	0.02	1.78	0.31
TF		0	0	0	0	0	0	100	0.049	23.33	59.56

Table 2: Distribution of $\hat{q} - q$ for data generated according to (2.1) with the noise term ε_t being i.i.d. $\mathcal{N}(0, 1)$ for various choices of f_t and σ_t given in Appendix A and competing methods introduced in Section 4. Also, the average Mean-Square Error of the resulting estimate of the signal f_t , average Hausdorff distance d_H given by (4.3) and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q} - q = 0$ or smallest average d_H and those within 10% of the highest, or, respectively, within 10% of the lowest.

Method	Model	$\hat{q} - q$							MSE	$d_H \times 10^2$	time
		≤ -3	-2	-1	0	1	2	≥ 3			
B&P	(M1)	82	9	2	7	0	0	0	0.832	14.15	0.26
e-cp3o		0	0	0	100	0	0	0	0.109	1.02	2.15
NMCD		0	0	0	98	2	0	0	0.149	1.43	1.28
NOT		0	0	0	99	1	0	0	0.112	1.05	0.08
NOT HT		0	0	0	97	3	0	0	0.127	1.35	0.09
NP-PELT		0	0	0	73	24	2	1	0.131	1.43	0.04
PELT		0	0	0	100	0	0	0	0.11	1.04	0
S3IB		0	0	0	94	5	1	0	0.113	1.17	0.11
SMUCE		0	1	15	84	0	0	0	0.192	2.23	0.23
WBS		0	0	0	98	2	0	0	0.11	1.05	0.11
B&P	(M2)	100	0	0	0	0	0	0	0.358	14.34	5.64
e-cp3o		100	0	0	0	0	0	0	0.142	8.12	194.18
NMCD		37	31	26	5	1	0	0	0.073	4.02	5.06
NOT		27	28	25	17	2	1	0	0.062	3.48	0.11
NOT HT		42	27	23	7	1	0	0	0.076	4.23	0.23
NP-PELT		1	12	26	25	17	16	3	0.067	3.91	0.54
PELT		92	7	0	1	0	0	0	0.106	7.28	0.01
S3IB		35	23	24	17	0	1	0	0.065	3.94	0.53
SMUCE		100	0	0	0	0	0	0	0.139	5.72	0.04
WBS		30	26	27	16	1	0	0	0.064	3.64	0.22
B&P	(M3)	0	0	100	0	0	0	0	0.246	3.94	146.74
NOT		0	0	0	99	1	0	0	0.032	1.47	0.54
TF		0	0	0	0	0	0	100	0.032	8.42	63.71
B&P	(M4)	16	55	28	1	0	0	0	0.336	6.48	167.31
NOT		0	0	0	98	2	0	0	0.039	2.08	0.47
TF		0	0	0	0	0	0	100	0.031	4.44	64.41
B&P	(M5)	0	0	8	92	0	0	0	0.044	3.31	380.84
NOT		0	0	5	93	2	0	0	0.045	3.52	0.48
TF		0	0	0	0	0	0	100	0.041	5.89	78.46
e-cp3o	(M6)	95	2	0	3	0	0	0	0.372	16.55	11.67
NMCD		0	0	15	79	6	0	0	0.058	3.35	4.78
NOT		0	0	10	89	1	0	0	0.045	2.07	1.22
NP-PELT		0	0	0	22	24	22	32	0.12	2.97	0.61
PELT		11	15	28	44	2	0	0	0.075	7.83	0.02
B&P	(M7)	0	0	35	65	0	0	0	0.066	6.47	44.26
NOT		0	1	37	62	0	0	0	0.064	5.78	0.31
TF		0	0	0	0	0	1	99	0.075	22.71	60.17

Table 3: Distribution of $\hat{q} - q$ for data generated according to (2.1) with the noise term ε_t being i.i.d. $\mathcal{N}(0, 2)$ for various choices of f_t and σ_t given in Appendix A and competing methods introduced in Section 4. Also, the average Mean-Square Error of the resulting estimate of the signal f_t , average Hausdorff distance d_H given by (4.3) and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q} - q = 0$ or smallest average d_H and those within 10% of the highest, or, respectively, within 10% of the lowest.

Method	Model	$\hat{q} - q$							MSE	$d_H \times 10^2$	time
		≤ -3	-2	-1	0	1	2	≥ 3			
B&P	(M1)	76	4	1	19	0	0	0	0.745	13.04	0.25
e-cp3o		0	0	0	100	0	0	0	0.097	0.87	2.13
NMCD		0	0	0	94	6	0	0	0.141	1.35	1.28
NOT		0	1	0	95	3	1	0	0.107	1.19	0.08
NOT HT		0	0	0	99	0	1	0	0.093	0.79	0.09
NP-PELT		0	0	0	71	22	6	1	0.141	1.57	0.04
PELT		0	0	0	69	13	14	4	0.145	1.4	0
S3IB		0	1	0	76	10	9	4	0.136	1.47	0.11
SMUCE		0	0	1	52	23	14	10	0.155	2.6	0.21
WBS		0	0	0	64	4	23	9	0.151	1.91	0.11
B&P	(M2)	100	0	0	0	0	0	0	0.311	12.55	5.36
e-cp3o		100	0	0	0	0	0	0	0.147	9.1	191.73
NMCD		15	36	37	12	0	0	0	0.06	3.37	5.06
NOT		51	21	17	9	2	0	0	0.079	4.8	0.11
NOT HT		23	26	36	15	0	0	0	0.054	3.08	0.23
NP-PELT		0	4	10	19	27	19	21	0.077	4.03	0.51
PELT		20	21	19	14	14	6	6	0.108	5.02	0.01
S3IB		88	8	2	2	0	0	0	0.13	10.22	0.5
SMUCE		14	16	23	22	6	8	11	0.108	6.02	0.03
WBS		21	12	12	15	15	10	15	0.104	4.98	0.22
B&P	(M3)	0	0	100	0	0	0	0	0.261	4.16	147.23
NOT		0	0	1	96	1	1	1	0.037	1.89	0.52
TF		0	0	0	0	0	0	100	0.035	8.42	64.08
B&P	(M4)	16	44	37	3	0	0	0	0.323	6.27	171.88
NOT		0	0	0	96	3	1	0	0.042	2.24	0.44
TF		0	0	0	0	0	0	100	0.032	4.38	66.53
B&P	(M5)	0	1	6	93	0	0	0	0.045	3.44	384.72
NOT		0	1	2	90	3	3	1	0.047	3.48	0.5
TF		0	0	0	0	0	0	100	0.041	5.91	78.1
e-cp3o	(M6)	96	3	1	0	0	0	0	0.481	17.95	11.91
NMCD		1	28	38	30	2	0	1	0.098	9.45	4.83
NOT		1	10	42	35	9	1	2	0.188	8.17	1.24
NP-PELT		0	1	4	14	22	16	43	0.359	5.34	0.75
PELT		22	22	35	17	3	1	0	0.215	12.8	0.03
B&P	(M7)	0	0	41	59	0	0	0	0.066	5.93	44.19
NOT		0	2	51	44	2	1	0	0.077	7.7	0.32
TF		0	0	0	0	0	0	100	0.075	22.42	60.33

Table 4: Distribution of $\hat{q} - q$ for data generated according to (2.1) with the noise term ε_t being i.i.d. Laplace $(0, (\sqrt{2})^{-1})$ (N.B. $\text{Var}(\varepsilon_t) = 1$ here) for various choices of f_t and σ_t given in Appendix A and competing methods introduced in Section 4. Also, the average Mean-Square Error of the resulting estimate of the signal f_t , average Hausdorff distance d_H given by (4.3) and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q} - q = 0$ or smallest average d_H and those within 10% of the highest, or, respectively, within 10% of the lowest.

Method	Model	$\hat{q} - q$							MSE	$d_H \times 10^2$	time
		≤ -3	-2	-1	0	1	2	≥ 3			
B&P	(M1)	65	12	0	23	0	0	0	0.67	10.76	0.26
e-cp3o		0	0	0	100	0	0	0	0.044	0.39	2.22
NMCD		0	0	0	94	6	0	0	0.092	0.81	1.31
NOT		0	0	0	94	5	1	0	0.046	0.57	0.08
NOT HT		0	0	0	98	2	0	0	0.045	0.47	0.1
NP-PELT		0	0	0	73	14	11	2	0.082	1.37	0.03
PELT		0	0	0	63	6	16	15	0.092	1.68	0
S3IB		0	0	0	54	7	20	19	0.096	1.84	0.11
SMUCE		0	0	0	45	22	19	14	0.091	2.53	0.21
WBS		0	0	0	44	3	28	25	0.105	2.44	0.11
B&P	(M2)	100	0	0	0	0	0	0	0.302	11.98	4.28
e-cp3o		100	0	0	0	0	0	0	0.126	5.87	197.26
NMCD		0	4	66	29	0	1	0	0.032	1.92	5.13
NOT		2	16	33	31	14	3	1	0.032	4.09	0.11
NOT HT		1	7	62	28	2	0	0	0.027	1.9	0.23
NP-PELT		0	0	6	22	20	23	29	0.048	3.91	0.46
PELT		0	3	16	19	20	12	30	0.066	3.98	0.01
S3IB		29	10	26	20	4	11	0	0.065	4.38	0.49
SMUCE		0	5	11	25	14	13	32	0.056	5.36	0.03
WBS		0	3	15	11	21	15	35	0.067	4.7	0.22
B&P	(M3)	0	0	100	0	0	0	0	0.217	3.63	149.51
NOT		0	0	0	99	1	0	0	0.015	1	0.63
TF		0	0	0	0	0	0	100	0.017	8.4	66.66
B&P	(M4)	0	0	10	90	0	0	0	0.081	2.78	175.34
NOT		0	0	0	94	5	1	0	0.019	1.51	0.54
TF		0	0	0	0	0	0	100	0.017	4.44	68.33
B&P	(M5)	0	0	0	100	0	0	0	0.019	2.29	392
NOT		0	0	0	96	4	0	0	0.019	2.33	0.53
TF		0	0	0	0	0	0	100	0.026	6.01	80.41
e-cp3o	(M6)	91	2	2	4	0	1	0	0.327	14.05	11.51
NMCD		0	12	47	36	5	0	0	0.053	8.56	4.94
NOT		0	4	17	35	25	12	7	0.08	6.1	1.26
NP-PELT		0	0	2	9	22	19	48	0.205	5.1	0.66
PELT		7	14	26	33	15	5	0	0.112	8.88	0.03
B&P	(M7)	0	0	0	99	1	0	0	0.021	2.5	45.59
NOT		0	0	8	79	11	2	0	0.03	4.28	0.32
TF		0	0	0	0	0	0	100	0.05	23.32	62.79

Table 5: Distribution of $\hat{q} - q$ for data generated according to (2.1) with the noise term ε_t being i.i.d. $(3/5)^{1/2}t_5$ (N.B. $\text{Var}(\varepsilon_t) = 1$ here) for various choices of f_t and σ_t given in Section A and competing methods introduced in Appendix 4. Also, the average Mean-Square Error of the resulting estimate of the signal f_t , average Hausdorff distance d_H given by (4.3) and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q} - q = 0$ or smallest average d_H and those within 10% of the highest, or, respectively, within 10% of the lowest.

NP-PELT overestimates q more frequently than in the $\mathcal{N}(0, 1)$ case. In the heavy-tailed scenarios ($\varepsilon_t \sim (3/5)^{1/2}t_5$ and $\varepsilon_t \sim \text{Laplace}(0, (\sqrt{2})^{-1})$), NOT, NOT HT, NMCD and e-cp3o, offer the best performance, while the other methods but B&P tend to slightly overestimate q .

For the **blocks** signal with $\mathcal{N}(0, 1)$ noise, WBS performs the best, S3IB is the second best, while NOT is the third best method, which can be seen from the corresponding values of the Hausdorff distance d_H and MSE. B&P, e-cp3o and SMUCE underestimate, while NP-PELT tends to overestimate the number of change-points. In the $\mathcal{N}(0, 2)$ case, NOT performs the best in terms of d_H and MSE, while WBS is the second best. In the heavy-tailed noise cases, performance of NOT HT and NMCD stands out, with the former achieving the best d_H and MSE, while PELT, NP-PELT, SMUCE tend to overestimate q .

Overall, we observe that only three methods, namely NMCD, NOT and NOT HT, perform reasonably well across all the examples with a piecewise constant signal.

- (S2) Two signals follow this scenario: (M3) **wave1** and (M4) **wave2**. For the **wave1** signal, we observe a pattern common across all considered scenarios for ε_t : typically B&P underestimates the number of changes in the slope coefficient, TF largely overestimates q while NOT tends to find the correct number of the change-points. The NOT estimates lie close to the true locations of the change-points, which can be seen from very low values of d_H . Moreover, NOT estimates of the underlying signal yields MSEs comparable to or even lower than the corresponding values for TF, despite the latter procedure having been designed solely for the estimation of f_t .

In (M4), NOT performs the best across all scenarios for ε_t , most often identifying the correct number of change-points. In the case of $\varepsilon_t \sim \mathcal{N}(0, 1)$ and $\varepsilon_t \sim (3/5)^{1/2}t_5$ B&P performs reasonably well, while in the remaining two scenarios it frequently fails to identify some of the change-points.

Finally, the NOT estimates are orders of magnitude quicker to compute than the competing estimators.

- (S3) The (M5) **mix** signal follows this scenario. In the case of $\varepsilon_t \sim \mathcal{N}(0, 1)$, NOT performs slightly better than B&P, always correctly identifying the number of change-points. TF performs well in terms of the average MSE, but it largely overestimates the number of change-points. On the other hand, NOT identifies the correct number of change-points more frequently than B&P when the noise $\varepsilon_t \sim \mathcal{N}(0, 2)$, but B&P achieves a slightly lower d_H in that scenario. In the heavy-tailed examples, B&P performs very well, while NOT slightly overestimates the number of change-points. However, we emphasise again that NOT is much quicker to compute than the competing methods.
- (S4) The (M6) **vo1** signal follows this scenario. In the cases of $\varepsilon_t \sim \mathcal{N}(0, 1)$ and $\varepsilon_t \sim \mathcal{N}(0, 2)$, NOT most frequently estimates the number of change-points correctly and achieves the lowest average d_H , while NMCD is the second best. In the heavy-tailed scenarios, NP-PELT achieves the best d_H , but it exhibits an overall tendency of overestimating the number of change-points. Besides, e-c3po and PELT in all cases underestimate q .

(S5) The (M7) **quad** signal follows this scenario. In the case of $\varepsilon_t \sim \mathcal{N}(0, 1)$, both NOT and B&P always correctly estimate the number of change-points, however, NOT estimates are on average closer to the true locations. The problem becomes more challenging for $\varepsilon_t \sim \mathcal{N}(0, 2)$, where all methods frequently fail to identify one change-point, with NOT being marginally better than B&P and significantly better than TF. The challenge here is that the signal between $t = 251$ to $t = 1000$ can be approximated by a quadratic function reasonably well, therefore SIC and other criteria may prefer a simpler model without a change-point at $t = 500$ when the standard deviation of the noise is relatively large. In the heavy-tailed cases, NOT slightly overestimates the number of change-points, however its performance in terms of d_H remains reasonably close to the performance of B&P, which is the best in these examples.

In all simulated scenarios, NOT is always either the best or not far from the best method. Importantly, it is quick to compute, which gives it a particular advantage over its competitors in Scenarios (S2), (S3) and (S5), where the computational complexity of the competing methods is polynomial, which is prohibitive for large sample sizes. Furthermore, NOT with the contrast function derived under the assumption that the noise is Gaussian is relatively robust against the misspecification in the distribution of ε_t .

5 Real data analysis

We present applications of the NOT methodology to three real data sets: oil price log-returns, temperature anomalies data and the UK House Price Index. All R code used in this section is available from our GitHub repository (Baranowski et al., 2016a).

5.1 OPEC Reference Basket oil price

NOT	NMCD	Event
29 April 2003	N/A	Invasion of Iraq
1 September 2008	28 August 2008	critical stage of the subprime mortgage crisis
27 January 2009	22 January 2009	tensions in the Gaza Strip
1 October 2009	23 October 2009	
12 November 2012	12 October 2012	beginning of a period of low volatility
30 September 2014	1 October 2014	
5 January 2016	21 January 2016	beginning of a sell-off leading the price to 12-year low
N/A	22 February 2016	

Table 6: Change-points detected using NOTWBS and NMCD methods in the daily OPEC Reference Basket oil price data from 1 January 2003 to 15 July 2016, with the majority of them dated.

We perform change-point analysis on the daily Organisation of the Petroleum Exporting Countries (OPEC) Reference Basket oil price from 1 January, 2003 to 15 July, 2016. The data were obtained from the OPEC database through the R package **Quandl** (McTaggart et al., 2016). Instead of working with the raw price series, we analyse the log-returns series $Y_t =$

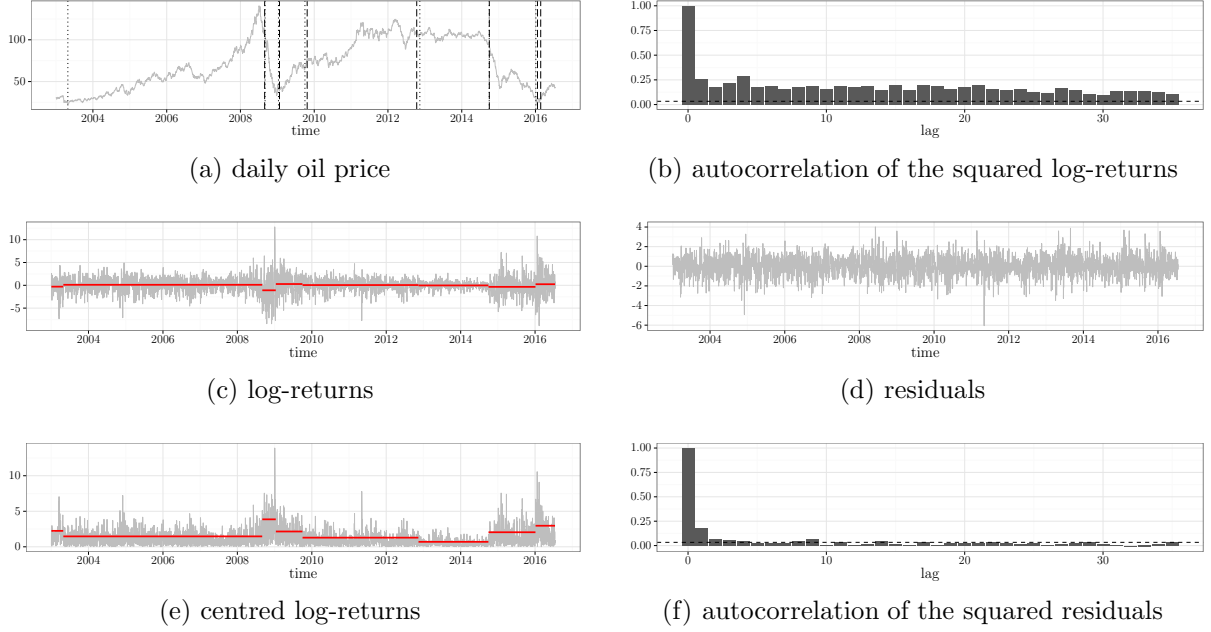


Figure 7: Change-point analysis on the daily OPEC Reference Basket oil price in USD from 1 January, 2003 to 15 July, 2016. Figure 7(a): price series P_t (thin grey), locations of the change-points detected with NOT (vertical dotted lines) and NMCD (vertical dashed lines). Figure 7(b): autocorrelation function of Y_t^2 . Figure 7(c): log-returns $Y_t = 100 \log(P_t/P_{t-1})$ (thin grey), the fitted piecewise-constant mean \hat{f}_t (thick red). Figure 7(d) residuals $\hat{\varepsilon}_t = (Y_t - \hat{f}_t)/\hat{\sigma}_t$. Figure 7(e): the centred log-returns $|Y_t - \hat{f}_t|$ (thin grey), fitted piecewise-constant volatility $\hat{\sigma}_t$ (thick red). Figure 7(f): autocorrelation of $\hat{\varepsilon}_t^2$. The exact locations of the change-points detected with NOT are given in Table 6.

$100 \log(P_t/P_{t-1})$, where P_t denotes the daily oil price. One of the stylised facts of the financial time series data is that the autocorrelation of assets returns are weak, while squared returns tend to exhibit strong autocorrelation, which is the case for the oil price time series (see Figure 7(b)). This phenomenon can be possibly explained by the existence of the structural breaks in the mean and variance structure of the data series (Mikosch and Stărică, 2004; Fryzlewicz et al., 2006). In this study, we apply NOT with the contrast function given by (2.10), which is designed to detect changes in both the mean and the volatility. For comparison, we also report change-points detected with the NMCD method of Zou et al. (2014), which was the second best method for change-point detection in Scenario (S4) in the simulation study of Section 4.

We apply Algorithm 2 to compute the NOT solution path and choose the model achieving the lowest SIC given by (3.1), setting the number of intervals drawn $M = 10000$ and the maximum number of change-points $q_{max} = 25$. Computations for the solution path and model selection are performed using the R package **not** (Baranowski et al., 2016b). For the NMCD procedure, we use the **nmcd** routine from the R package **nmcdR** (Zou and Lancezhang, 2014), setting the maximum number of change-points to $q_{max} = 25$ as well.

Figure 7 illustrates the results of our analysis. The oil price time series and the locations

of the change-points identified by NOT and NMCD can be seen in Figure 7(a). Both methods discover 7 change-points, largely agreeing on their locations, in the sense that for 6 out of 7 NOT estimates, NMCD detects a change-point nearby. However, NMCD does not indicate any change-point around the first change-point identified by NOT on 29 April 2003. This date can be clearly related to the end of the 2003 invasion of Iraq, which initiated the upward trend in the oil price lasting almost ceaselessly until the beginning of the 2008–09 financial crisis. On the other hand, NMCD indicates two change-points in the first quarter of 2016, while NOT finds a single change-point in that period. Table 6 lists the exact locations of the change-points detected by the two methods and the events that can be related to some of them. Figure 7(f) shows the autocorrelation function for the squared residuals obtained by subtracting the sample mean and dividing by the standard deviations from the data in each segment. It appears that there is little autocorrelation in the squares of the residuals, meaning that (S4) models the data in this example reasonably well.

5.2 Temperature anomalies

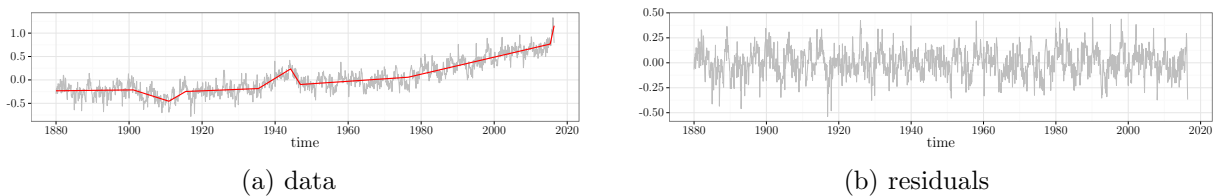


Figure 8: Change-point analysis for the GISSTEMP data set introduced in Section 5.2. Figure 8(a): the data series Y_t (thin grey) and \hat{f}_t estimated using change-points returned by NOT (thick red). Figure 8(b): residuals $\hat{\varepsilon}_t = Y_t - \hat{f}_t$.

For the second application, we analyse the GISS Surface Temperature anomalies data set available from GISTEMP Team (2016), consisting of monthly temperature anomalies recorded from January 1880 to June 2016. The anomaly here is defined as the difference between the average global temperature in a given month and the baseline value, being the average calculated for that time of the year over the 30-year period from 1951 to 1980; for more details see Hansen et al. (2010). This and similar anomalies series are frequently studied in literature with a particular focus on identifying change-points in the data, see e.g. Ruggieri (2013) or James and Matteson (2015).

The plot of the data (Figure 8(a)) clearly indicates the presence of a linear trend with several change-points in the temperature anomalies series. The corresponding changes are not abrupt, therefore we believe that Scenario (S2) with change-points in the slope of the trend is most appropriate here. To detect the locations of the change-points, we apply Algorithm 2 with the contrast given by (2.7), combined with the SIC criterion to determine the best model on the solution path. The maximum number of change-points for NOT is set to $q_{max} = 25$ and $M = 50000$.

Figure 8 shows the data, the NOT estimate of the piecewise-linear trend and the empirical residuals. We identify 8 change-points located at the following dates: March 1901, December 1910, July 1915, June 1935, April 1944, December 1946, June 1976 and May 2015. Previous

studies conducted on similar temperature anomalies series (observed at a yearly frequency and obtained from a different source), report change-points around 1910, 1945 and 1976 (see Ruggieri (2013) for an overview of a number of related analyses). In addition to the change-points around these dates, NOT identifies two periods, 1901–1915 and 1935–1946, where local deviations from the baseline trend are clearly visible. We also observe a long-lasting upward trend in the anomalies series starting in December 1946. NOT estimates indicate that the slope of the trend is increasing, with the most recent change-point in May 2015.

5.3 UK House Price Index

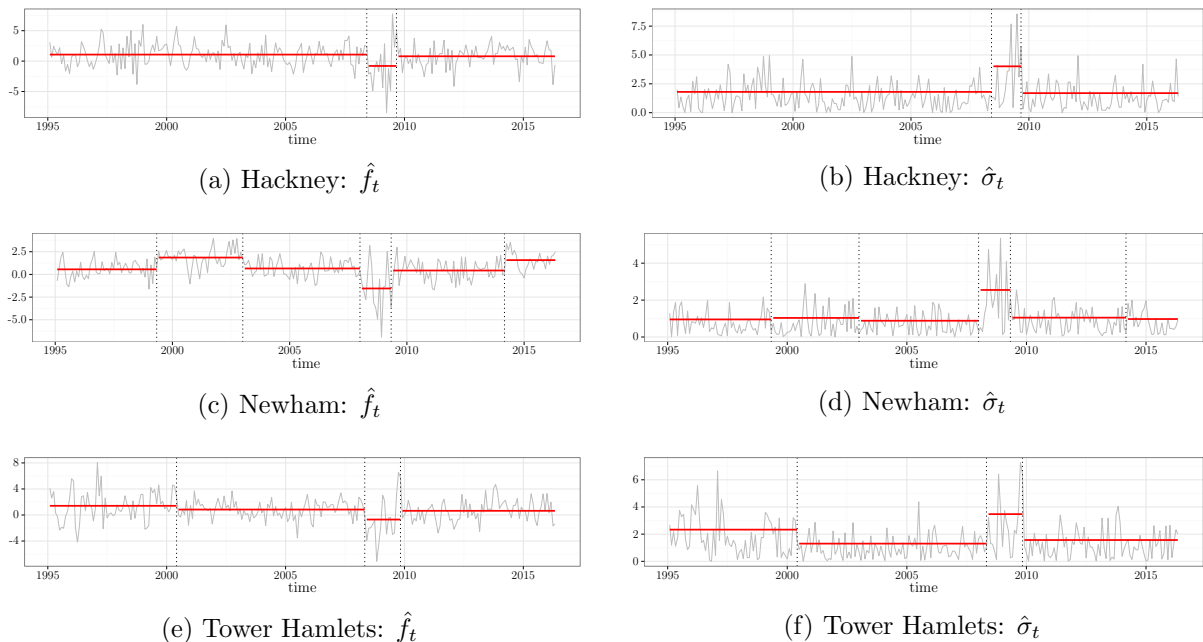


Figure 9: Change-point analysis for the monthly percentage changes in the UK House Price Index from January 1995 to May 2016. Figure 9(a), 9(c) and 9(e): the monthly percentage changes Y_t and the fitted piecewise-constant mean \hat{f}_t , between the change-points estimated with NOT. Figure 9(b), 9(d) and 9(f): $|Y_t - \hat{f}_t|$ and the fitted piecewise-constant standard deviation $\hat{\sigma}_t$, between the change-points estimated with NOT.

In our final example, we analyse monthly percentage changes in the UK House Price Index (HPI) which provides an overall estimate of the changes in house prices across the UK. The data and a detailed description of how the index is calculated are available online from UK Land Registry (2016). Fryzlewicz (2016), who proposed a method for signal estimation and change-point detection in Scenario (S1), used this data set to illustrate the performance of his methodology. We perform similar analysis, assuming the more flexible Scenario (S4), allowing for changes both in the mean and the variance of the series, which, we argue, leads to some additional insights and better-interpretable estimates in this case.

As in Fryzlewicz (2016), we analyse the percentage changes in the HPI for three London boroughs, namely Hackney, Newham and Tower Hamlets, all of which are located in East

London. Hackney and Tower of Hamlets border on the City of London, a major business and financial district, with the latter being a home to Canary Wharf, another important financial centre. On the other hand, Newham, located to the east of Hackney and Tower Hamlets, hosted the London 2012 Olympic Games which involved large-scale investment in that borough.

Figure 9 shows monthly percentage changes in HPI for the analysed boroughs and the corresponding NOT estimates, obtained using the contrast function (2.10). As recommended in Section 3.4, we set the number of intervals drawn in the procedure to $M = 10000$ and choose the threshold that minimises the SIC criterion (3.1). For better comparability, NOT is applied with the same random seed for each data series.

In contrast to Fryzlewicz (2016), whose TGUH method estimates at least 10 change-points in each HPI series, we detect just a few change-points in the data, facilitating the interpretation of the results. Furthermore, for all three boroughs, NOT estimates two change-points (one around March 2008 and one around September 2009) that can clearly be linked to the 2008–2009 financial crisis and the concurrent collapse of the housing market. Estimated standard deviations for that period are much larger than the estimates corresponding to the other segments of piecewise-constancy, suggesting that in this example Scenario (S4) may be more relevant than (S1) considered in Fryzlewicz (2016). It is also interesting to observe that, with the exception of Tower Hamlets from January 1995 to April 2000 and the 2008–2009 financial crisis for all boroughs, the estimated standard deviations oscillate around a baseline level (different for each series).

The period of a larger volatility for Tower Hamlets in Figure 9(f), observed from January 1995 to April 2000, can possibly be explained by developments in Canary Wharf, which in the past was a dock complex closed in 1980. Gordon (2001) claims that the project of converting Canary Wharf into a business district “was politically controversial and widely regarded as a planning disaster” which “(in 1992) failed as a result of six factors: a recession in the London property market, competition from the City of London, poor transport links, few British tenants, complicated finances and developer overconfidence”. Over the 1995–2000 period, the situation in the London property reversed, which combined with a development of new public transport lines in Canary Wharf led to the success of the project. According to Gordon (2001), “when the Jubilee underground line opened in 2000, Canary Wharfs resurrection was complete”.

Finally, it is interesting to observe that over two periods, namely March 1991 to November 2002 and January 2014 to May 2016, the HPI for Newham (Figure 9(c)) was increasing at a rate higher than for the other two boroughs.

Acknowledgements

Piotr Fryzlewicz’s work was supported by the Engineering and Physical Sciences Research Council grant no. EP/L014246/1.

A Simulation models

- (M1) **teeth**: piecewise-constant f_t (in Scenario (S1)), $T = 512$, $q = 7$ change-points at $\tau = 64, 128, \dots, 448$, with the corresponding jump sizes $-2, 2, -2, \dots, -2$, starting intercept $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \dots, T$.
- (M2) **blocks**: piecewise-constant f_t (in Scenario (S1)), $T = 2024$, $q = 11$ change-points at $\tau = 205, 267, 308, 472, 512, 820, 902, 1332, 1557, 1598, 1659$, with the corresponding jump sizes $1.464, -1.830, 1.098, -1.464, 1.830, -1.537, 0.768, 1.574, -1.135, 0.769, -1.537$, starting intercept $f_1 = 0$, $\sigma_t = 1$ for $t = 1, \dots, T$. This signal is widely analysed in the literature, see e.g. Fryzlewicz (2014).
- (M3) **wave1**: piecewise-linear f_t without jumps in the intercept (in Scenario (S2)), $T = 1408$, $q = 7$ change-points at $\tau = 256, 512, 768, 1024, 1152, 1280, 1344$, with the corresponding changes in slopes $1 \cdot 2^{-6}, -2 \cdot 2^{-6}, 3 \cdot 2^{-6}, \dots, -7 \cdot 2^{-6}$, starting intercept $f_1 = 1$ and slope $f_2 - f_1 = 2^{-8}$, $\sigma_t = 1$ for $t = 1, \dots, T$.
- (M4) **wave2**: piecewise-linear f_t without jumps in the intercept (in Scenario (S2)), $T = 1500$, $q = 9$ change-points at $\tau = 150, 300, \dots, 1350$, with the corresponding changes in slopes $2^{-5}, -2^{-5}, 2^{-5}, \dots, -2^{-5}$, starting intercept $f_1 = 2^{-1}$ and slope $f_2 - f_1 = 2^{-6}$, $\sigma_t = 1$ for $t = 1, \dots, T$.
- (M5) **mix**: piecewise-linear f_t with jumps in the intercept (in Scenario (S3)), length $T = 2048$, $q = 7$ change-points at $\tau = 256, 512, \dots, 1792$, with the corresponding changes in the intercept $0, -1, 0, 0, 2, -1, 0$ and in the slope $2^{-6}, -2^{-6}, -2^{-6}, 2^{-6}, 0, 2^{-6}, -2^{-5}$, starting value for the intercept $f_1 = 0$ and slope $f_2 - f_1 = 0$, $\sigma_t = 1$ for $t = 1, \dots, T$.
- (M6) **vol**: piecewise-constant f_t and σ_t (in Scenario (S4)), $T = 2048$, $q = 7$ changes at $\tau = 256, 512, \dots, 1792$ with the corresponding jumps in f_t and σ_t being $1, 0, -2, 0, 2, -1, 0$ and $0, 1, 0, 1, 0, -1, 1$, respectively, initial values $f_1 = \sigma_1 = 1$.
- (M7) **quad**: piecewise-quadratic f_t (in Scenario (S5)), $T = 1000$, $q = 3$ change-points at $\tau = 100, 250, 500$, with the corresponding changes in the intercept $2, -2, 0$, in the slope $0, -10^{-1}, 10^{-1}$ and in the quadratic coefficient $0, 0, 2 \times 10^{-5}$, the initial values $f_1 = f_2 - f_1 = f_3 - 2f_2 + f_1 = 0$, $\sigma_t = 1$ for all $t = 1, \dots, T$.

B Proofs

B.1 Some useful lemmas

B.1.1 The piecewise constant case

Lemma 1. *Let $g(x, y) = \frac{xy}{x+y}$ and suppose that $\min(x, y) > 0$. Then*

$$g(x, y) \geq \frac{1}{2} \min(x, y).$$

Proof. Without loss of generality, assume that $x \geq y$. Then $g(x, y) \geq \frac{xy}{2x} \geq y/2 = \min(x, y)/2$. \square

Lemma 2. Suppose $\mathbf{f} = (f_1, \dots, f_T)'$ is piecewise-constant vector as in Scenario (S1), and τ_1, \dots, τ_q are the locations of the change-points. Suppose $1 \leq s < e \leq T$, such that $\tau_{j-1} < s \leq \tau_j < e \leq \tau_{j+1}$ for some $j = 1, \dots, q$. Let $\eta = \min\{\tau_j - s + 1, e - \tau_j\}$ and $\Delta_j^{\mathbf{f}} = |f_{\tau_{j+1}} - f_{\tau_j}|$. Then

$$\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}) = \max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) \begin{cases} \geq \frac{1}{\sqrt{2}} \eta^{1/2} \Delta_j^{\mathbf{f}}, \\ \leq \eta^{1/2} \Delta_j^{\mathbf{f}}. \end{cases}$$

Proof. For any $s \leq b < e$, by simple algebra, we have

$$\mathcal{C}_{s,e}^b(\mathbf{f}) = \begin{cases} \sqrt{\frac{b-s+1}{l(e-b)}}(e - \tau_j)|f_{\tau_{j+1}} - f_{\tau_j}|, & b \leq \tau_j; \\ \sqrt{\frac{(\tau_j-s+1)(e-\tau_j)}{l}}|f_{\tau_{j+1}} - f_{\tau_j}|, & b = \tau_j; \\ \sqrt{\frac{e-b}{l(b-s+1)}}(\tau_j - s + 1)|f_{\tau_{j+1}} - f_{\tau_j}|, & b \geq \tau_j, \end{cases} \quad (\text{B.1})$$

where $l = s - e + 1$. Now $\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}) = \max_{s \leq b \leq e} \mathcal{C}_{s,e}^b(\mathbf{f})$ follows from the fact that $\mathcal{C}_{s,e}^b(\mathbf{f})$ is increasing (as a function of b) for $1 \leq b \leq \tau_j$ and decreasing for $\tau_j \leq b \leq e$. To prove the lower bound, we set $\eta_L = \tau_j - s + 1$ and $\eta_R = e - \tau_j$ and observe that $\eta_L \geq \eta$ and $\eta_R \geq \eta$. Therefore by Lemma 1, $\frac{\eta_L \eta_R}{\eta_L + \eta_R} \geq \frac{\eta}{2}$. Noting that $l = \eta_L + \eta_R$ we bound

$$\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}) = \sqrt{\frac{(\tau_j - s + 1)(e - \tau_j)}{l}}|f_{\tau_{j+1}} - f_{\tau_j}| \begin{cases} \geq (\eta/2)^{1/2} \Delta_j^{\mathbf{f}}, \\ \leq \eta^{1/2} \Delta_j^{\mathbf{f}}. \end{cases}$$

which completes the proof. \square

Lemma 3. Suppose $\mathbf{f} = (f_1, \dots, f_T)'$ is piecewise-constant vector as in Scenario (S1), and τ_1, \dots, τ_q are the locations of the change-points. Suppose $1 \leq s < e \leq T$ such that $\tau_{j-1} < s \leq \tau_j$ and $\tau_{j+1} < e \leq \tau_{j+2}$ for some $j = 1, \dots, q-1$. Then

$$\max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) \leq (\tau_j - s + 1)^{1/2} \Delta_j^{\mathbf{f}} + (e - \tau_{j+1})^{1/2} \Delta_{j+1}^{\mathbf{f}}$$

where $\Delta_j^{\mathbf{f}} = |f_{\tau_{j+1}} - f_{\tau_j}|$.

Proof. Suppose that $b^* = \arg\max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f})$. Then

$$\begin{aligned} 0 &\leq \|\mathbf{f} - \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^{b^*} \rangle \boldsymbol{\psi}_{s,e}^{b^*} - \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle \mathbf{1}_{s,e}\|^2 = \|\mathbf{f} - \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle \mathbf{1}_{s,e}\|^2 - \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^{b^*} \rangle^2 \\ &\leq \|\mathbf{f} - f_{\tau_{j+1}} \sqrt{s - e + 1} \mathbf{1}_{s,e}\|^2 - \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^{b^*} \rangle^2 \\ &= (\tau_j - s + 1)(\Delta_j^{\mathbf{f}})^2 + (e - \tau_{j+1})(\Delta_{j+1}^{\mathbf{f}})^2 - \left(\max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) \right)^2. \end{aligned}$$

It then follows that

$$\max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) \leq \sqrt{(\tau_j - s + 1)(\Delta_j^{\mathbf{f}})^2 + (e - \tau_{j+1})(\Delta_{j+1}^{\mathbf{f}})^2} \leq (\tau_j - s + 1)^{1/2} \Delta_j^{\mathbf{f}} + (e - \tau_{j+1})^{1/2} \Delta_{j+1}^{\mathbf{f}}.$$

\square

Lemma 4. Suppose $\mathbf{f} = (f_1, \dots, f_T)'$ is piecewise-constant vector as in Scenario (S1). Pick any interval $[s, e] \subset [1, T]$ such that $[s, e - 1]$ contains exactly one change-point τ_j . Let $\rho = |\tau_j - b|$, $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$, $\eta_L = \tau_j - s + 1$ and $\eta_R = e - \tau_j$. Then,

$$\|\psi_{s,e}^b \langle \mathbf{f}, \psi_{s,e}^b \rangle - \psi_{s,e}^{\tau_j} \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle\|_2^2 = (\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s,e}^b(\mathbf{f}))^2.$$

Moreover,

1. for any $\tau_j \leq b < e$, $(\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s,e}^b(\mathbf{f}))^2 = \frac{\rho \eta_L}{\rho + \eta_L} (\Delta_j^{\mathbf{f}})^2$;
2. for any $s \leq b < \tau_j$, $(\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s,e}^b(\mathbf{f}))^2 = \frac{\rho \eta_R}{\rho + \eta_R} (\Delta_j^{\mathbf{f}})^2$.

Proof. First, we note that since there is only one change-point in $[s, e - 1]$, the restriction of \mathbf{f} on $[s, e]$, i.e. $\mathbf{f}|_{[s,e]} = (0, \dots, 0, f_s, \dots, f_e, 0, \dots, 0)'$ can be decomposed into

$$\mathbf{f}|_{[s,e]} = \psi_{s,e}^{\tau_j} \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle + \mathbf{1}_{s,e} \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle,$$

where we also used the fact that $\psi_{s,e}^{\tau_j}$ and $\mathbf{1}_{s,e}$ are orthonormal. Note that $\psi_{s,e}^b$ and $\mathbf{1}_{s,e}$ are also orthonormal, it follows that

$$\langle \mathbf{f}, \psi_{s,e}^b \rangle = \langle \mathbf{f}|_{[s,e]}, \psi_{s,e}^b \rangle = \langle \psi_{s,e}^{\tau_j} \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle + \mathbf{1}_{s,e} \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle, \psi_{s,e}^b \rangle = \langle \psi_{s,e}^{\tau_j}, \psi_{s,e}^b \rangle \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle.$$

Therefore,

$$\langle \mathbf{f}, \psi_{s,e}^b \rangle^2 = \langle \mathbf{f}, \psi_{s,e}^b \rangle \langle \psi_{s,e}^{\tau_j}, \psi_{s,e}^b \rangle \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle,$$

and thus

$$\begin{aligned} \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle^2 - \langle \mathbf{f}, \psi_{s,e}^b \rangle^2 &= \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle^2 + \langle \mathbf{f}, \psi_{s,e}^b \rangle^2 - 2 \langle \mathbf{f}, \psi_{s,e}^b \rangle \langle \psi_{s,e}^{\tau_j}, \psi_{s,e}^b \rangle \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle \\ &= \|\psi_{s,e}^b \langle \mathbf{f}, \psi_{s,e}^b \rangle - \psi_{s,e}^{\tau_j} \langle \mathbf{f}, \psi_{s,e}^{\tau_j} \rangle\|_2^2. \end{aligned}$$

Here in the above final step, we used the fact that $\|\psi_{s,e}^{\tau_j}\|_2^2 = \|\psi_{s,e}^b\|_2^2 = 1$.

Second, for the sake of brevity, we only prove the case of $b \geq \tau_j$. Let $l = e - s + 1$, $x = b - s + 1$, and thus $\rho = x - \eta_L$. Using (B.1), we get

$$\begin{aligned} (\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s,e}^b(\mathbf{f}))^2 &= \left(\frac{\eta_L(l - \eta_L)}{l} - \frac{\eta_L^2(l - x)}{lx} \right) |f_{\tau_j+1} - f_{\tau_j}|^2 \\ &= \frac{\eta_L(x - \eta_L)}{x} (\Delta_j^{\mathbf{f}})^2 = \left(\frac{\rho \eta_L}{\eta_L + \rho} \right) (\Delta_j^{\mathbf{f}})^2. \end{aligned}$$

□

B.1.2 The piecewise linear continuous case

Lemma 5. Suppose $\mathbf{f} = (f_1, \dots, f_T)'$ is piecewise-linear vector as in Scenario (S2), and τ_1, \dots, τ_q are the locations of the change-points. Suppose $1 \leq s < e \leq T$, such that $\tau_{j-1} \leq s < \tau_j < e \leq \tau_{j+1}$ for some $j = 1, \dots, q$. Let $\eta = \min\{\tau_j - s, e - \tau_j\}$ and $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$. Then

$$\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}) = \max_{s < b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) \begin{cases} \geq \frac{1}{\sqrt{24}} \eta^{3/2} \Delta_j^{\mathbf{f}}, \\ \leq \frac{1}{\sqrt{3}} (\eta + 1)^{3/2} \Delta_j^{\mathbf{f}}. \end{cases}$$

Proof. First, we show that $\mathcal{C}_{s,e}^b(\mathbf{f})$ is maximised at $b = \tau_j$. Using the notation from the proof of Lemma 4, we have that

$$\mathbf{f}|_{[s,e]} = \phi_{s,e}^{\tau_j} \langle \mathbf{f}, \phi_{s,e}^{\tau_j} \rangle + \gamma_{s,e} \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle + \mathbf{1}_{s,e} \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle.$$

Therefore, it follows that

$$\|\mathbf{f}|_{[s,e]}\|_2^2 = \langle \mathbf{f}, \phi_{s,e}^{\tau_j} \rangle^2 + \langle \mathbf{f}, \gamma_{s,e} \rangle^2 + \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle^2. \quad (\text{B.2})$$

For any $b \in \{s+1, \dots, \tau_j-1, \tau_j+1, \dots, e-1\}$, it is clear that $\mathbf{f}|_{[s,e]}$ does not lie in the span of $\phi_{s,e}^b$, $\gamma_{s,e}$ and $\mathbf{1}_{s,e}$. Consequently, by projecting $\mathbf{f}|_{[s,e]}$ onto these three bases, we have that

$$\|\mathbf{f}|_{[s,e]}\|^2 > \langle \mathbf{f}, \phi_{s,e}^b \rangle^2 + \langle \mathbf{f}, \gamma_{s,e} \rangle^2 + \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle^2. \quad (\text{B.3})$$

Comparing (B.3) with (B.2) entails that $|\langle \mathbf{f}, \phi_{s,e}^{\tau_j} \rangle| > |\langle \mathbf{f}, \phi_{s,e}^b \rangle|$ for any $b \neq \tau_j$.

Secondly, set $\eta_L = \tau_j - s$ and $\eta_R = e - \tau_j$. After some calculation, we get that

$$\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}) = \left\{ \frac{\eta_L(\eta_L + 1)\eta_R(\eta_R + 1)(2\eta_L\eta_R + \eta_L + \eta_R + 2)}{6l(l^2 - 1)} \right\} \Delta_j^{\mathbf{f}},$$

where $l = e - s + 1$. Also, we have $\eta_L \geq \eta$, $\eta_R \geq \eta$ and $l = \eta_L + \eta_R + 1$. To prove the lower bound, we observe that

$$\begin{aligned} & \left\{ \frac{\eta_L(\eta_L + 1)\eta_R(\eta_R + 1)(2\eta_L\eta_R + \eta_L + \eta_R + 2)}{6l(l^2 - 1)} \right\} \\ & \geq \left\{ \frac{1}{6} \frac{(\eta_L + 1)\eta_R}{l} \frac{\eta_L(\eta_R + 1)}{l} \frac{2 \min(\eta_L, \eta_R) \{\max(\eta_L, \eta_R) + 1\}}{l} \right\} \geq \left\{ \frac{\eta^3}{24} \right\}, \end{aligned}$$

where the last inequality is obtained applying Lemma 1 three times. For the upper bound, we notice that $2\eta_L\eta_R + \eta_L + \eta_R + 2 \leq 2(\eta_L + 1)(\eta_R + 1)$ which implies

$$\left\{ \frac{\eta_L(\eta_L + 1)\eta_R(\eta_R + 1)(2\eta_L\eta_R + \eta_L + \eta_R + 2)}{6l(l^2 - 1)} \right\} \leq \left\{ \frac{1}{3} \frac{\eta_L\eta_R(\eta_L + 1)^2(\eta_R + 1)^2}{(l - 1)l^2} \right\} \leq \left\{ \frac{(\eta + 1)^3}{3} \right\}.$$

□

Lemma 6. Suppose $\mathbf{f} = (f_1, \dots, f_T)'$ is piecewise-linear vector as in Scenario (S2), and τ_1, \dots, τ_q are the locations of the change-points. Suppose $1 \leq s < e \leq T$ such that $\tau_{j-1} \leq s \leq \tau_j$ and $\tau_{j+1} \leq e \leq \tau_{j+2}$ for some $j = 1, \dots, q-1$. Then

$$\max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) \leq \frac{1}{\sqrt{3}}(\tau_j - s + 1)^{3/2} \Delta_j^{\mathbf{f}} + \frac{1}{\sqrt{3}}(e - \tau_{j+1} + 1)^{3/2} \Delta_{j+1}^{\mathbf{f}},$$

where $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_{j-1}} - f_{\tau_{j+1}}|$.

Proof. Suppose that $b^* = \arg\max_{s \leq b \leq e} \mathcal{C}_{s,e}^b(\mathbf{f})$. Then

$$\begin{aligned} 0 & \leq \|\mathbf{f} - \langle \mathbf{f}, \phi_{s,e}^{b^*} \rangle \phi_{s,e}^{b^*} - \langle \mathbf{f}, \gamma_{s,e} \rangle \gamma_{s,e} - \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle \mathbf{1}_{s,e}\|^2 = \|\mathbf{f} - \langle \mathbf{f}, \gamma_{s,e} \rangle \gamma_{s,e} - \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle \mathbf{1}_{s,e}\|^2 - \langle \mathbf{f}, \phi_{s,e}^{b^*} \rangle^2 \\ & = \frac{1}{6}(\tau_j - s)(\tau_j - s + 1)(2\tau_j - 2s + 1)(\Delta_j^{\mathbf{f}})^2 + \frac{1}{6}(e - \tau_{j+1})(e - \tau_{j+1} + 1)(2e - 2\tau_{j+1} + 1)(\Delta_{j+1}^{\mathbf{f}})^2 \\ & \quad - \left(\max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) \right)^2. \end{aligned}$$

It then follows that

$$\begin{aligned} \max_{s \leq b < e} \mathcal{C}_{s,e}^b(\mathbf{f}) &\leq \{(\tau_j - s + 1)^3(\Delta_j^{\mathbf{f}})^2/3 + (e - \tau_{j+1} + 1)^3(\Delta_{j+1}^{\mathbf{f}})^2/3\} \\ &\leq \frac{1}{\sqrt{3}}(\tau_j - s + 1)^{3/2}\Delta_j^{\mathbf{f}} + \frac{1}{\sqrt{3}}(e - \tau_{j+1} + 1)^{3/2}\Delta_{j+1}^{\mathbf{f}}. \end{aligned}$$

□

Lemma 7. Suppose $\mathbf{f} = (f_1, \dots, f_T)'$ is piecewise-linear vector as in Scenario (S2), and τ_1, \dots, τ_q are the locations of the change-points. Suppose $1 \leq s < e \leq T$, such that $\tau_{j-1} \leq s < \tau_j < e \leq \tau_{j+1}$ for some $j = 1, \dots, q$. Let $\rho = |\tau_j - b|$, $\eta_L = \tau_j - s$, $\eta_R = e - \tau_j$ and $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_{j-1}} - f_{\tau_{j+1}}|$. Then,

$$\|\phi_{s,e}^b \langle \mathbf{f}, \phi_{s,e}^b \rangle - \phi_{s,e}^{\tau_j} \langle \mathbf{f}, \phi_{s,e}^{\tau_j} \rangle\|_2^2 = (\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s,e}^b(\mathbf{f}))^2. \quad (\text{B.4})$$

Moreover,

1. for any $\tau_j \leq b < e$, $(\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s,e}^b(\mathbf{f}))^2 \geq \frac{1}{63} \min(\rho, \eta_L)^3 (\Delta_j^{\mathbf{f}})^2$;
2. for any $s < b \leq \tau_j$, $(\mathcal{C}_{s,e}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s,e}^b(\mathbf{f}))^2 \geq \frac{1}{63} \min(\rho, \eta_R)^3 (\Delta_j^{\mathbf{f}})^2$.

Proof. The proof of (B.4) is very similar to that shown in Lemma 4, so is omitted for brevity. In the following, we only deal with the case of $\tau_j \leq b < e$. Note that

$$\begin{aligned} \|\phi_{s,e}^b \langle \mathbf{f}, \phi_{s,e}^b \rangle - \phi_{s,e}^{\tau_j} \langle \mathbf{f}, \phi_{s,e}^{\tau_j} \rangle\|_2^2 &= \|\phi_{s,e}^b \langle \mathbf{f}, \phi_{s,e}^b \rangle + \gamma_{s,e} \langle \mathbf{f}, \gamma_{s,e} \rangle + \mathbf{1}_{s,e} \langle \mathbf{f}, \mathbf{1}_{s,e} \rangle - \mathbf{f}|_{[s,e]}\|_2^2 \\ &\geq \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{f}|_{[s,b]} - a_0 \mathbf{1}_{s,b} - a_1 \gamma_{s,b}\|_2^2 + \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{f}|_{[b+1,e]} - a_0 \mathbf{1}_{b+1,e} - a_1 \gamma_{b+1,e}\|_2^2 \\ &\geq \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{f}|_{[s,b]} - a_0 \mathbf{1}_{s,b} - a_1 \gamma_{s,b}\|_2^2. \end{aligned}$$

Recalling the definitions of $\alpha_{s,b}^{\tau_j}$ and $\beta_{s,b}^{\tau_j}$ in (2.6), and writing $d = b - s + 1$. After some calculations (similar to what has already been carried out in deriving $\phi_{s,e}^b$), we obtain that

$$\begin{aligned} \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{f}|_{[s,b]} - a_0 \mathbf{1}_{s,b} - a_1 \gamma_{s,b}\|_2^2 &= \left[(3\eta_L + \rho + 2) \alpha_{s,b}^{\tau_j} \beta_{s,b}^{\tau_j} + (3\rho + \eta_L + 2) \alpha_{s,b}^{\tau_j} (\beta_{s,b}^{\tau_j})^{-1} \right]^{-2} (\Delta_j^{\mathbf{f}})^2 \\ &= \frac{1}{6} (\Delta_j^{\mathbf{f}})^2 d(d^2 - 1) [1 + \rho\eta_L + (\rho + 1)(\eta_L + 1)] \times \\ &\quad \left[(d + 2\eta_L + 1)^2 \frac{\rho(\rho + 1)}{\eta_L(\eta_L + 1)} + (d + 2\rho + 1)^2 \frac{\eta_L(\eta_L + 1)}{\rho(\rho + 1)} + 2(d + 2\eta_L + 1)(d + 2\rho + 1) \right]^{-1}. \end{aligned}$$

Notice that the above equation is symmetric with respect to η_L and ρ . Without loss of generality, here we proceed by assuming that $\eta_L \geq \rho$. Since $(d + 2\eta_L + 1) + (d + 2\rho + 1) = 4d$, it follows that $(d + 2\eta_L + 1)(d + 2\rho + 1) \leq 4d^2$. Therefore,

$$\begin{aligned} \min_{a_0, a_1 \in \mathbb{R}} \|\mathbf{f}|_{[s,b]} - a_0 \mathbf{1}_{s,b} - a_1 \gamma_{s,b}\|_2^2 &\geq \frac{1}{6} (\Delta_j^{\mathbf{f}})^2 d(d^2 - 1) [2(\eta_L + 1)\rho] \left[(3d)^2 + (2d)^2 \frac{(\eta_L + 1)^2}{\rho^2} + 8d^2 \right]^{-1} \\ &\geq \frac{1}{6} (\Delta_j^{\mathbf{f}})^2 d^2 (d - 1) [2(\eta_L + 1)\rho] \left[21d^2 \frac{(\eta_L + 1)^2}{\rho^2} \right]^{-1} \geq \frac{1}{63} \rho^3 (\Delta_j^{\mathbf{f}})^2, \end{aligned}$$

where in the last step, we used the fact that $\frac{d-1}{\eta_L+1} \geq 1$ for $\rho \geq 1$ (and note that the last above-displayed equation also holds if $\rho = 0$).

Finally, we remark that the case of $s < b \leq \tau_j$ can also be handled by symmetry. \square

B.2 Proof of Theorem 1

Here we informally discuss our proof strategy, which could be generalised to other scenarios. Intuitively speaking, lemmas from Appendix B.1 deal with noiseless versions of the change-point estimation problems. In order to apply these results to show the consistency of estimated number of change-points, we need to control $\|\mathcal{C}_{s,e}^b(\mathbf{Y}) - \mathcal{C}_{s,e}^b(\mathbf{f})\|$ for every (s, e, b) , which can be achieved using Bonferroni in Step 1. Note that for any fixed interval with start-point s and end-point e , to decide whether b_1 or b_2 is a more suitable change-point candidate inside this interval, we only need to look at the value of $\mathcal{C}_{s,e}^{b_1}(\mathbf{Y}) - \mathcal{C}_{s,e}^{b_2}(\mathbf{Y})$. Therefore, when establishing the convergence rate of the estimated change-point location, we control the distance between $\mathcal{C}_{s,e}^{b_1}(\mathbf{Y}) - \mathcal{C}_{s,e}^{b_2}(\mathbf{Y})$ and its noiseless analogue $\mathcal{C}_{s,e}^{b_1}(\mathbf{f}) - \mathcal{C}_{s,e}^{b_2}(\mathbf{f})$ (after proper normalisation) for all tuples (s, e, b_1, b_2) in Step 2. In Step 3, we show that given a properly chosen threshold and a large enough M , both bounds in Step 1 and Step 2 hold, and for each change-point τ_j , there exists an interval from F_T^M that contains only this change-point and both its start- and end- points are sufficiently far away from other change-points. Since we are dealing with the narrowest-over-threshold intervals, the actual intervals that our NOT algorithm pick must have length no longer than the ones we considered in Step 3, thus could only contain precisely one change-point. So in Step 4, it suffices to investigate a single change-point detection problem, where we can use lemmas from Appendix B.1 and the bound in Step 2 to establish the convergence rate for its location estimation. Finally, in Step 5, we show that after detecting all the change-points, the NOT algorithm stops with no further detection. This is because the remaining elements $[s, e] \in F_T^M$ to be considered either have no change-point inside, or have one/two change-points that are very close to its start- or/and end- points, thus their corresponding $\max_b \mathcal{C}_{s,e}^b(\mathbf{Y})$ cannot exceed the given threshold in views of the property of its noiseless analogue and the bound from Step 1.

Now we proceed to the technical details.

Proof. We shall prove the following more specific result, which in turn implies (2.11).

$$\mathbb{P}\left(\hat{q} = q, \max_{j=1,\dots,q} \left(|\hat{\tau}_j - \tau_j|(\Delta_j^{\mathbf{f}})^2\right) \leq C_3 \log T\right) \geq 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/36)^M, \quad (\text{B.5})$$

Step One.

Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_T)'$ and $\lambda_T = \sqrt{8 \log T}$. Define the set

$$A_T = \left\{ \max_{s,b,e: 1 \leq s \leq b < e \leq T} |\mathcal{C}_{s,e}^b(\boldsymbol{\varepsilon})| \leq \lambda_T \right\}.$$

Note that for any $1 \leq s \leq b < e \leq T$, $\mathcal{C}_{s,e}^b(\boldsymbol{\varepsilon})$ follows a standard normal distribution. Therefore, using the Bonferroni bound, we get

$$\mathbb{P}(A_T^c) \leq \frac{T^3}{6} \frac{2e^{-(\sqrt{8 \log T})^2/2}}{\sqrt{8 \log T} \sqrt{2\pi}} \leq \frac{T^{-1}}{12\sqrt{\pi}}.$$

Moreover, because $\mathcal{C}_{s,e}^b(\mathbf{Y}) - \mathcal{C}_{s,e}^b(\mathbf{f}) = \mathcal{C}_{s,e}^b(\boldsymbol{\varepsilon})$, so A_T also implies that

$$\left\{ \max_{s,b,e: 1 \leq s \leq b < e \leq T} |\mathcal{C}_{s,e}^b(\mathbf{Y}) - \mathcal{C}_{s,e}^b(\mathbf{f})| \leq \lambda_T \right\}.$$

Step Two.

Define the set

$$B_T = \left\{ \max_{j=1,\dots,q} \max_{\substack{\tau_{j-1} < s \leq \tau_j \\ \tau_j < e \leq \tau_{j+1} \\ s \leq b < e}} \frac{\left| \left\langle \boldsymbol{\psi}_{s,e}^b \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^b \rangle - \boldsymbol{\psi}_{s,e}^{\tau_j} \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^{\tau_j} \rangle, \boldsymbol{\varepsilon} \right\rangle \right|}{\left\| \boldsymbol{\psi}_{s,e}^b \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^b \rangle - \boldsymbol{\psi}_{s,e}^{\tau_j} \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^{\tau_j} \rangle \right\|_2} \leq \lambda_T \right\}.$$

Again, for any $1 \leq s \leq b < e \leq T$, $\frac{|\langle \boldsymbol{\psi}_{s,e}^b \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^b \rangle - \boldsymbol{\psi}_{s,e}^{\tau_j} \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^{\tau_j} \rangle, \boldsymbol{\varepsilon} \rangle|}{\left\| \boldsymbol{\psi}_{s,e}^b \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^b \rangle - \boldsymbol{\psi}_{s,e}^{\tau_j} \langle \mathbf{f}, \boldsymbol{\psi}_{s,e}^{\tau_j} \rangle \right\|_2}$ follows a standard normal distribution, so using a similar argument, we get

$$\mathbb{P}(B_T^c) \leq \frac{T^{-1}}{12\sqrt{\pi}}.$$

Step Three.

To fix the ideas, for $j = 1, \dots, q$, we define intervals

$$\mathcal{I}_j^L = (\tau_j - \delta_T/3, \tau_j - \delta_T/6) \quad (\text{B.6})$$

$$\mathcal{I}_j^R = (\tau_j + \delta_T/6, \tau_j + \delta_T/3) \quad (\text{B.7})$$

Note that these intervals all contain at least one integer as long as $\delta_T > 6$. This is always true for sufficiently large T , as it follows from Conditions 1 and 2 that $\delta_T > \underline{C} \log T / \underline{f}$. Recall that F_T^M is the set of M randomly drawn intervals with endpoints in $\{1, \dots, T\}$. Denote by $[s_1, e_1], \dots, [s_M, e_M]$ the elements of F_T^M and let

$$D_T^M = \left\{ \forall j = 1, \dots, q, \exists k \in \{1, \dots, M\}, \text{ s.t. } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R \right\}. \quad (\text{B.8})$$

We have that

$$\begin{aligned} \mathbb{P}((D_T^M)^c) &\leq \sum_{j=1}^q \Pi_{m=1}^M \left(1 - \mathbb{P}(s_m \times e_m \in \mathcal{I}_j^L \times \mathcal{I}_j^R) \right) \\ &\leq q \left(1 - \frac{\delta_T^2}{6^2 T^2} \right)^M \leq \frac{T}{\delta_T} \left(1 - \frac{\delta_T^2}{36 T^2} \right)^M. \end{aligned}$$

Therefore, $\mathbb{P}(A_T \cap B_T \cap D_T^M) \geq 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/36)^M$.

In the rest of the proof, we assume that A_T, B_T and D_T^M all hold. We give the constants as follows:

$$C_1 = 2\sqrt{C_3} + \sqrt{8}, \quad C_2 = \frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{\underline{C}}, \quad C_3 = 32\sqrt{2} + 48.$$

These constants could be further refined by applying the Bonferroni bound more carefully. But since our main aim is to establish the rate, we chose not to pursue this direction further. In addition, here we need to make sure that $\underline{C}C_2 > C_1$, and thus $C_2\delta_T^{1/2}\underline{f}_T > C_1\sqrt{\log T}$, i.e. we can select $\zeta_T \in [C_1\sqrt{\log T}, C_2\delta_T^{1/2}\underline{f}_T]$. This is indeed the case because \underline{C} is sufficiently large.

Step Four.

We focus on a generic interval $[s, e]$ such that

$$\exists j \in \{1, \dots, q\}, \exists k \in \{1, \dots, M\}, \text{ s.t. } [s_k, e_k] \subset [s, e] \text{ and } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R \quad (\text{B.9})$$

Fix such an interval $[s, e]$ and let $j \in \{1, \dots, q\}$ and $k \in \{1, \dots, M\}$ be such that (B.9) is satisfied. Let $b_k^* = \operatorname{argmax}_{s_k \leq b \leq e_k} \mathcal{C}_{s_k, e_k}^b(\mathbf{Y})$. By construction, $[s_k, e_k]$ satisfies $\tau_j - s_k + 1 \geq \delta_T/6$ and $e_k - \tau_j > \delta_T/6$. Denote by

$$\begin{aligned} \mathcal{M}_{s,e} &= \{m : [s_m, e_m] \in F_T^M, [s_m, e_m] \subset [s, e]\}; \\ \mathcal{O}_{s,e} &= \{m \in \mathcal{M}_{s,e} : \max_{s_m \leq b < e_m} \mathcal{C}_{s_m, e_m}^b(\mathbf{Y}) > \zeta_T\} \end{aligned}$$

Our first aim is to show that $\mathcal{O}_{s,e}$ is non-empty. This follows from Lemma 2 and the calculation below.

$$\begin{aligned} \mathcal{C}_{s_k, e_k}^{b_k^*}(\mathbf{Y}) &\geq \mathcal{C}_{s_k, e_k}^{\tau_j}(\mathbf{Y}) \\ &\geq \mathcal{C}_{s_k, e_k}^{b_k^*}(\mathbf{f}) - \lambda_T \geq \left(\frac{\delta_T}{6}\right)^{1/2} |f_{\tau_j+1} - f_{\tau_j}| - \lambda_T \geq \left(\frac{\delta_T}{6}\right)^{1/2} \underline{f}_T - \lambda_T \\ &= \left(\frac{1}{\sqrt{6}} - \frac{\lambda_T}{\delta_T^{1/2} \underline{f}_T}\right) \delta_T^{1/2} \underline{f}_T \geq \left(\frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{\underline{C}}\right) \delta_T^{1/2} \underline{f}_T = C_2 \delta_T^{1/2} \underline{f}_T > \zeta_T. \end{aligned}$$

Let $m^* = \operatorname{argmin}_{m \in \mathcal{O}_{s,e}} (e_m - s_m + 1)$ and $b^* = \operatorname{argmax}_{s_{m^*} \leq b < e_{m^*}} \mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{Y})$. Observe that $[s_{m^*}, e_{m^*})$ must contain at least one change-point. Indeed, if that was not the case, we would have $\mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{f}) = 0$ and

$$\mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{Y}) = |\mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{Y}) - \mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{f})| \leq \lambda_T \leq \zeta_T$$

which contradicts $\mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{Y}) > \zeta_T$. On the other hand, $[s_{m^*}, e_{m^*})$ cannot contain more than one change-points, because $e_{m^*} - s_{m^*} + 1 \leq e_k - s_k + 1 \leq \delta_T$, as we picked the *narrowest*-over-threshold interval.

Without loss of generality, assume $\tau_j \in [s_{m^*}, e_{m^*}]$. Denote by $\eta_L = \tau_j - s_{m^*} + 1$, $\eta_R = e_{m^*} - \tau_j$ and $\eta_T = (C_1 - \sqrt{8})^2 (\Delta_j^{\mathbf{f}})^{-2} \log T$, where $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$. We claim that $\min(\eta_L, \eta_R) > \eta_T$, because $\min(\eta_L, \eta_R) \leq \eta_T$ and Lemma 2 result in

$$\begin{aligned} \mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{Y}) &\leq \mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{f}) + \lambda_T \leq \mathcal{C}_{s_{m^*}, e_{m^*}}^{\tau_j}(\mathbf{f}) + \lambda_T \leq \eta_T^{1/2} \Delta_j^{\mathbf{f}} + \lambda_T \\ &= (C_1 - \sqrt{8} + \sqrt{8}) \sqrt{\log T} = C_1 \sqrt{\log T} \leq \zeta_T, \end{aligned}$$

which contradicts $\mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{Y}) > \zeta_T$.

We are now in the position to prove $|b^* - \tau_j| \leq C_3 \log T / (\Delta_j^f)^2$. The arguments we use here are simpler and slightly more general than Lemma A.3 of Fryzlewicz (2014). Our aim is to find ϵ_T such that for any $b \in \{s_{m^*}, s_{m^*} + 1, \dots, e_{m^*} - 1\}$ with $|b - \tau_j| > \epsilon_T$, we always have

$$(\mathcal{C}_{s_{m^*}, e_{m^*}}^{\tau_j}(\mathbf{Y}))^2 - (\mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{Y}))^2 > 0. \quad (\text{B.10})$$

This would then imply that $|b^* - \tau_j| \leq \epsilon_T$. By expansion and rearranging the terms (using the fact that $f_t = Y_t + \varepsilon_t$), we see that (B.10) is equivalent to

$$\begin{aligned} \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \rangle^2 - \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^b \rangle^2 &> \langle \varepsilon, \psi_{s_{m^*}, e_{m^*}}^b \rangle^2 - \langle \varepsilon, \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \rangle^2 \\ &+ 2 \left\langle \varepsilon, \psi_{s_{m^*}, e_{m^*}}^b \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^b \rangle - \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \rangle \right\rangle. \end{aligned} \quad (\text{B.11})$$

In the following, we assume that $b \geq \tau_j$. The case that $b < \tau_j$ can be handled in a similar fashion. By Lemma 4, we have

$$\langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \rangle^2 - \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^b \rangle^2 = (\mathcal{C}_{s^*, e^*}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{f}))^2 = \frac{|b - \tau_j| \eta_L}{|b - \tau_j| + \eta_L} (\Delta_j^f)^2 := \kappa.$$

In addition, since A_T and B_T hold, we have that

$$\begin{aligned} \langle \varepsilon, \psi_{s_{m^*}, e_{m^*}}^b \rangle^2 - \langle \varepsilon, \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \rangle^2 &\leq \lambda_T^2, \\ 2 \left\langle \varepsilon, \psi_{s_{m^*}, e_{m^*}}^b \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^b \rangle - \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \rangle \right\rangle \\ &\leq 2 \|\psi_{s_{m^*}, e_{m^*}}^b \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^b \rangle - \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \langle \mathbf{f}, \psi_{s_{m^*}, e_{m^*}}^{\tau_j} \rangle\|_2 \lambda_T = 2\kappa^{1/2} \lambda_T, \end{aligned}$$

where the last equality also comes from Lemma 4. Consequently, (B.11) can be deduced from the stronger inequality $\kappa - 2\lambda_T \kappa^{1/2} - \lambda_T^2 > 0$. This quadratic inequality is implied by $\kappa > (\sqrt{2} + 1)^2 \lambda_T^2$, and could be restricted further to

$$\frac{2|b - \tau_j| \eta_L}{|b - \tau_j| + \eta_L} \geq \min(|b - \tau_j|, \eta_L) > (32\sqrt{2} + 48)(\Delta_j^f)^{-2} \log T = C_3 (\Delta_j^f)^{-2} \log T. \quad (\text{B.12})$$

But since

$$\eta_L \geq \eta_T = (C_1 - \sqrt{8})^2 (\Delta_j^f)^{-2} \log T = (2\sqrt{C_3})^2 (\Delta_j^f)^{-2} \log T > C_3 (\Delta_j^f)^{-2} \log T,$$

we see that (B.12) is equivalent to $|b - \tau_j| > C_3 (\Delta_j^f)^{-2} \log T$. To sum up, $|b^* - \tau_j| (\Delta_j^f)^2 > C_3 \log T$ would result in (B.10), a contradiction. So we have proved that $|b^* - \tau_j| (\Delta_j^f)^2 \leq C_3 \log T$.

Step Five.

Using the arguments given above which are valid on the event $A_T \cap B_T \cap D_T^M$, we can now proceed with the proof of the theorem as follows. At the start of Algorithm 1 we have $s = 1$ and $e = T$ and, provided that $q \geq 1$, condition (B.9) is satisfied. Therefore the

algorithm detects a change-point b^* in that interval such that $|b^* - \tau_j| \leq C_3 \log T (\Delta_j^f)^{-2}$. By construction, we also have that $|b^* - \tau_j| < 2/3\delta_T$. This in turn implies that for all $l = 1, \dots, q$ such that $\tau_l \in [s, e]$ and $l \neq j$ we have either $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [s, b^*]$ or $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [b^* + 1, e]$. Therefore (B.9) is satisfied within each segment containing at least one change-point. Note that before all q change-points are detected, each change-point will not be detected twice. To see this, we suppose that τ_j has already been detected by b , then for all intervals $[s_k, e_k] \subset [\tau_j - C_3 \log T (\Delta_j^f)^{-2} + 1, \tau_j - C_3 \log T (\Delta_j^f)^{-2} + 2/3\delta_T + 1] \cup [\tau_j + C_3 \log T (\Delta_j^f)^{-2} - 2/3\delta_T, \tau_j + C_3 \log T (\Delta_j^f)^{-2}]$, Lemma 2, together with the event A_T , guarantees that

$$\max_{s_k \leq b < e_k} \mathcal{C}_{s_k, e_k}^b(\mathbf{Y}) \leq \max_{s \leq b < e} \mathcal{C}_{s_k, e_k}^b(\mathbf{f}) + \sqrt{8 \log T} \leq \sqrt{C_3 \log T (\Delta_j^f)^{-2} \Delta_j^f} + \sqrt{8 \log T} \leq C_1 \sqrt{\log T} \leq \zeta_T.$$

Once all the change-points are detected, we then only need to consider $[s_k, e_k]$ such that

$$[s_k, e_k] \subset [\tau_j - C_3 \log T (\Delta_j^f)^{-2} + 1, \tau_{j+1} + C_3 \log T (\Delta_{j+1}^f)^{-2}]$$

for $j = 0, \dots, q$, where we set $\Delta_0^f = \Delta_{q+1}^f = \infty$ for notational convenience. It follows from Lemma 3 (within A_T) that

$$\begin{aligned} \max_{s_k \leq b < e} \mathcal{C}_{s_k, e_k}^b(\mathbf{Y}) &\leq \max_{s \leq b < e} \mathcal{C}_{s_k, e_k}^b(\mathbf{f}) + \sqrt{8 \log T} \\ &\leq \sqrt{C_3 \log T (\Delta_j^f)^{-2} \Delta_j^f} + \sqrt{C_3 \log T (\Delta_{j+1}^f)^{-2} \Delta_{j+1}^f} + \sqrt{8 \log T} \\ &< (2\sqrt{C_3} + \sqrt{8})\sqrt{\log T} = C_1 \sqrt{\log T} \leq \zeta_T. \end{aligned}$$

Hence the algorithm terminates and no further change-points are detected. \square

B.3 Proof of Theorem 2

Proof. The proof proceeds in analogy to the proof of Theorem 1. In five steps we shall establish the following result,

$$\mathbb{P} \left(\hat{q} = q, \max_{j=1, \dots, q} \left(|\hat{\tau}_j - \tau_j| (\Delta_j^f)^{2/3} \right) \leq C_3 (\log T)^{1/3} \right) \geq 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/36)^M, \quad (\text{B.13})$$

which in turn implies (2.12).

Step One and Step Two

We define the following two events

$$\begin{aligned} A_T &= \left\{ \max_{s, b, e: 1 \leq s \leq b < e \leq T} |\mathcal{C}_{s, e}^b(\boldsymbol{\epsilon})| \leq \lambda_T \right\}, \\ B_T &= \left\{ \max_{j=1, \dots, q} \max_{\substack{\tau_{j-1} < s \leq \tau_j \\ \tau_j < e \leq \tau_{j+1} \\ s \leq b < e}} \frac{\left| \langle \boldsymbol{\psi}_{s, e}^b \langle \mathbf{f}, \boldsymbol{\psi}_{s, e}^b \rangle - \boldsymbol{\psi}_{s, e}^{\tau_j} \langle \mathbf{f}, \boldsymbol{\psi}_{s, e}^{\tau_j} \rangle, \boldsymbol{\epsilon} \rangle \right|}{\| \boldsymbol{\psi}_{s, e}^b \langle \mathbf{f}, \boldsymbol{\psi}_{s, e}^b \rangle - \boldsymbol{\psi}_{s, e}^{\tau_j} \langle \mathbf{f}, \boldsymbol{\psi}_{s, e}^{\tau_j} \rangle \|_2} \leq \lambda_T \right\}, \end{aligned}$$

where $\lambda_T = \sqrt{8 \log T}$. Arguments as those used in Step One and Step Two of the proof of Theorem 2.11 show that $\mathbb{P}(A_T^c) \leq \frac{T^{-1}}{12\sqrt{\pi}}$ and $\mathbb{P}(B_T^c) \leq \frac{T^{-1}}{12\sqrt{\pi}}$.

Step Three

In the rest of the proof, we assume that A_T , B_T and D_T^M all hold, where the last event is given by (B.8). Exactly as in the proof of Theorem 2.11, we show that $\mathbb{P}(A_T \cap B_T \cap D_T^M) \geq 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/36)^M$.

We give the constants as follows:

$$C_1 = 2\sqrt{\frac{2}{3}}C_3^{3/2} + \sqrt{8}, \quad C_2 = \frac{1}{72} - \frac{2\sqrt{2}}{\underline{C}}, \quad C_3 = 2\sqrt[3]{7} \left(3(1 + \sqrt{2})\right)^{2/3}.$$

We require \underline{C} to be sufficiently large such that $\underline{C}C_2 > C_1$. Consequently it is possible to select $\zeta_T \in [C_1\sqrt{\log T}, C_2\delta_T^{3/2}\underline{f}_T]$.

Step Four

Consider a generic interval $[s, e]$ satisfying

$$\exists j \in \{1, \dots, q\}, \exists k \in \{1, \dots, M\}, \text{ s.t. } [s_k, e_k] \subset [s, e] \text{ and } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R \quad (\text{B.14})$$

and define events

$$\begin{aligned} \mathcal{M}_{s,e} &= \{m : [s_m, e_m] \in F_T^M, [s_m, e_m] \subset [s, e]\}, \\ \mathcal{O}_{s,e} &= \{m \in \mathcal{M}_{s,e} : \max_{s_m \leq b < e_m} \mathcal{C}_{s_m, e_m}^b(\mathbf{Y}) > \zeta_T\}. \end{aligned}$$

Let $b_k^* = \operatorname{argmax}_{s_k \leq b \leq e_k} \mathcal{C}_{s_k, e_k}^b(\mathbf{Y})$. We have

$$\begin{aligned} \mathcal{C}_{s_k, e_k}^{b_k^*}(\mathbf{Y}) &\geq \mathcal{C}_{s_k, e_k}^{\tau_j}(\mathbf{Y}) \\ &\geq \mathcal{C}_{s_k, e_k}^{b_k^*}(\mathbf{f}) - \lambda_T \geq \frac{1}{\sqrt{24}}(\delta_T/6)^{3/2} \Delta_j^{\mathbf{f}} - \lambda_T \geq \frac{1}{72}\delta_T^{3/2}\underline{f}_T - \lambda_T \\ &= \left(\frac{1}{72} - \frac{\lambda_T}{\delta_T^{3/2}\underline{f}_T}\right) \delta_T^{1/2}\underline{f}_T \geq \left(\frac{1}{72} - \frac{2\sqrt{2}}{\underline{C}}\right) \delta_T^{3/2}\underline{f}_T = C_2\delta_T^{3/2}\underline{f}_T > \zeta_T, \end{aligned}$$

where the third inequality above follows from Lemma 5, therefore $\mathcal{O}_{s,e}$ is non-empty.

Let $m^* = \operatorname{argmin}_{m \in \mathcal{O}_{s,e}} (e_m - s_m + 1)$ and $b^* = \operatorname{argmax}_{s_{m^*} \leq b < e_{m^*}} \mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{Y})$. Arguing exactly as in Step Four in the proof of Theorem 1, we show that $[s_{m^*}, e_{m^*})$ must contain exactly one change-point. Without loss of generality, assume that $\tau_j \in [s_{m^*}, e_{m^*})$. Let $\eta_L = \tau_j - s_{m^*}$, $\eta_R = e_{m^*} - \tau_j$ and $\eta_T = (\sqrt{3}(C_1 - \sqrt{8})\sqrt{\log T}(\Delta_j^{\mathbf{f}})^{-1})^{2/3} - 1$. We observe that $\min(\eta_L, \eta_R) > \eta_T$, as $\min(\eta_L, \eta_R) \leq \eta_T$ and Lemma 5 implies that

$$\begin{aligned} \mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{Y}) &\leq \mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{f}) + \lambda_T \leq \mathcal{C}_{s_{m^*}, e_{m^*}}^{\tau_j}(\mathbf{f}) + \lambda_T \leq \frac{1}{\sqrt{3}}(\eta_T + 1)^{3/2} \Delta_j^{\mathbf{f}} + \lambda_T \\ &= (C_1 - \sqrt{8} + \sqrt{8})\sqrt{\log T} = C_1\sqrt{\log T} \leq \zeta_T, \end{aligned}$$

contradicting $\mathcal{C}_{s_{m^*}, e_{m^*}}^{b^*}(\mathbf{Y}) > \zeta_T$.

We are now in position to prove that $|b^* - \tau_j| \leq C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} := \epsilon_T$. Let $b \in \{s_{m^*} + 1, \dots, e_{m^*} - 2\}$ and define $\kappa = ((\mathcal{C}_{s_k, e_k}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s_k, e_k}^b(\mathbf{f}))^2)$. We claim that

$$(\mathcal{C}_{s_{m^*}, e_{m^*}}^{\tau_j}(\mathbf{Y}))^2 - (\mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{Y}))^2 > 0, \quad (\text{B.15})$$

when $|b - \tau_j| > \epsilon_T$. Inequality (B.15) does not hold for $b = b^*$, so proving the claim suffices to demonstrate that $|b^* - \tau_j| \leq \epsilon_T$. Without loss of generality, we consider the case of $b > \tau_j$. Using arguments as those in Step Four of the proof of Theorem 1 we can show that (B.15) is implied by $\kappa > (\sqrt{2} + 1)^2 \lambda_T^2$, where $\kappa = (\mathcal{C}_{s_{m^*}, e_{m^*}}^{\tau_j}(\mathbf{f}))^2 - (\mathcal{C}_{s_{m^*}, e_{m^*}}^b(\mathbf{f}))^2$. By Lemma 7, $\kappa > (\sqrt{2} + 1)^2 \lambda_T^2$ is implied by

$$\min(|b - \tau_j|, \eta_L) > \left(63(\Delta_j^{\mathbf{f}})^{-2} \cdot 8(\sqrt{2} + 1)^2 \log T\right)^{1/3} = C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}$$

However, for sufficiently large T ,

$$\begin{aligned} \eta_L &> \eta_T = (\sqrt{3}(C_1 - \sqrt{8}))^{2/3}(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} - 1 > (C_1 - \sqrt{8})^{2/3}(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} \\ &= (C_3^{3/2} + \sqrt{8} - \sqrt{8})^{2/3}(\Delta_j^{\mathbf{f}})^{-2/3} = C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} = \epsilon_T, \end{aligned}$$

hence $|b - \tau_j| > \epsilon_T$ implies (B.15), so it must hold that $|b^* - \tau_j| \leq \epsilon_T$.

Step 5

Using the arguments given above which are valid on the event $A_T \cap B_T \cap D_T^M$, we can now proceed with the proof of the theorem as follows. At the start of Algorithm 1 we have $s = 1$ and $e = T$ and, provided that $q \geq 1$, condition (B.9) is satisfied. Therefore the algorithm detects a change-point b^* in that interval such that $|b^* - \tau_j| \leq C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}$. By construction, we also have that $|b^* - \tau_j| < 2/3\delta_T$. This in turn implies that for all $l = 1, \dots, q$ such that $\tau_l \in [s, e]$ and $l \neq j$ we have either $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [s, b^*]$ or $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [b^* + 1, e]$. Therefore (B.9) is satisfied within each segment containing at least one change-point. Note that before all q change-points are detected, each change-point will not be detected twice. To see this, we suppose that τ_j has already been detected by b , then for all intervals $[s_k, e_k] \subset [\tau_j - \epsilon_T + 1, \tau_j - \epsilon_T + 2/3\delta_T + 1] \cup [\tau_j + \epsilon_T - 2/3\delta_T, \tau_j + \epsilon_T]$, Lemma 5, together with the event A_T , guarantees that

$$\begin{aligned} \max_{s_k \leq b < e_k} \mathcal{C}_{s_k, e_k}^b(\mathbf{Y}) &\leq \max_{s \leq b < e} \mathcal{C}_{s_k, e_k}^b(\mathbf{f}) + \sqrt{8 \log T} \leq \frac{1}{\sqrt{3}}(C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} + 1)^{3/2} \Delta_j^{\mathbf{f}} + \sqrt{8 \log T} \\ &\leq (2\sqrt{\frac{2}{3}}C_3^{3/2} + \sqrt{8})\sqrt{\log T} = C_1\sqrt{\log T} \leq \zeta_T \end{aligned}$$

Once all the change-points are detected, we then only need to consider $[s_k, e_k]$ such that

$$[s_k, e_k] \subset [\tau_j - C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} + 1, \tau_{j+1} + C_3(\Delta_{j+1}^{\mathbf{f}})^{-2/3}(\log T)^{1/3}]$$

for $j = 0, \dots, q$, where we set $\Delta_0^f = \Delta_{q+1}^f = \infty$ for notational convenience. It follows from Lemma 6 (within A_T) that

$$\begin{aligned} \max_{s_k \leq b < e} \mathcal{C}_{s_k, e_k}^b(\mathbf{Y}) &\leq \max_{s \leq b < e} \mathcal{C}_{s_k, e_k}^b(\mathbf{f}) + \sqrt{8 \log T} \\ &\leq \frac{1}{\sqrt{3}} (C_3(\Delta_j^f)^{-2/3} (\log T)^{1/3})^{3/2} \Delta_j^f + \frac{1}{\sqrt{3}} (C_3(\Delta_j^f)^{-2/3} (\log T)^{1/3})^{3/2} \Delta_{j+1}^f + \sqrt{8 \log T} \\ &= \left(\frac{2}{\sqrt{3}} C_3^{3/2} + \sqrt{8} \right) \sqrt{\log T} \leq C_1 \sqrt{\log T} \leq \zeta_T. \end{aligned}$$

Hence the algorithm terminates and no further change-points are detected. \square

References

- A. Antoniadis, J. Bigot, and S. Lambert-Lacroix. Peaks detection and alignment for mass spectrometry data. *Journal de la Société Française de Statistique*, 151:17–37, 2010.
- J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66:47–78, 1998.
- J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18:1–22, 2003.
- R. Baranowski and P. Fryzlewicz. wbs: Wild binary segmentation for multiple change-point detection, 2015. URL <https://CRAN.R-project.org/package=wbs>. R package version 1.3.
- R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-over-threshold detection of multiple change-points and change-point-like features: Simulation code. <https://github.com/rbaranowski/not-num-ex>, 2016a.
- R. Baranowski, Y. Chen, and P. Fryzlewicz. not: Narrowest-over-threshold change-point detection, 2016b. URL <https://cran.r-project.org/web/packages/not>. R package version 1.0.
- N. Cahill, S. Rahmstorf, and A. C. Parnell. Change points of global temperature. *Environmental Research Letters*, 10:084002, 2015.
- H. P. Chan and G. Walther. Detection with the scan and the average likelihood ratio. *Statistica Sinica*, 23:409–428, 2013.
- A. Cleynen, G. Rigai, and M. Koskas. Segmentor3isback: A fast segmentation algorithm, 2013. URL <https://CRAN.R-project.org/package=Segmentor3IsBack>. R package version 1.8.
- L. Dagum and R. Menon. OpenMP: an industry standard API for shared-memory programming. *IEEE computational science and engineering*, 5:46–55, 1998.
- C. De Boor. *A Practical Guide to Splines*. Springer, 2001.

- R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.
- J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*. CRC Press, 1996.
- X. Fang, J. Li, and D. Siegmund. Segmentation and estimation of change-point models. *arXiv preprint arXiv:1608.03032*, 2016.
- K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society*, 76:495–580, 2014.
- P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42:2243–2281, 2014.
- P. Fryzlewicz. Tail-greedy bottom-up data decompositions and fast multiple change-point detection. 2016. URL <http://stats.lse.ac.uk/fryzlewicz/tguh/tguh.pdf>.
- P. Fryzlewicz, T. Sapatinas, and S. S. Rao. A Haar–Fisz technique for locally stationary volatility estimation. *Biometrika*, 93:687–704, 2006.
- GISTEMP Team. GISS Surface Temperature Analysis (GISTEMP). <http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/>, 2016. [Online; accessed 1-July-2016].
- D. L. Gordon. The resurrection of Canary Wharf. *Planning Theory and Practice*, 2:149–168, 2001.
- F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.
- J. Hansen, R. Ruedy, M. Sato, and K. Lo. Global surface temperature change. *Reviews of Geophysics*, 48:1–29, 2010.
- D. M. Hawkins. Fitting multiple change-point models to data. *Computational Statistics and Data Analysis*, 37:323–341, 2001.
- K. Haynes, P. Fearnhead, and I. A. Eckley. A computationally efficient nonparametric approach for changepoint detection. *arXiv preprint arXiv:1602.01254*, 2016a.
- K. Haynes, R. Killick, P. Fearnhead, and I. Eckley. changepoint.np: Methods for nonparametric changepoint detection, 2016b. URL <https://CRAN.R-project.org/package=changepoint.np>. R package version 0.0.2.
- T. Hotz and H. Sieling. stepR: Fitting step-functions, 2016. URL <http://CRAN.R-project.org/package=stepR>. R package version 1.0-4.
- N. A. James and D. S. Matteson. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62:1–25, 2014.

- N. A. James and D. S. Matteson. Change points via probabilistically pruned objectives. *arXiv preprint arXiv:1505.04302*, 2015.
- R. Killick and I. A. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58:1–19, 2014.
- R. Killick, P. Fearn, and I. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107:1590–1598, 2012a.
- R. Killick, C. Nam, J. Aston, and I. Eckley. The changepoint repository, 2012b. URL <http://changepoint.info/>.
- S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. L1 trend filtering. *SIAM Review*, 51:339–360, 2009.
- I. Koch. On the asymptotic performance of median smoothers in image analysis and non-parametric regression. *Annals of Statistics*, 24:1648–1666, 1996.
- E. D. Kolaczyk. Nonparametric estimation of gamma-ray burst intensities using Haar wavelets. *The Astrophysical Journal*, 483:340–349, 1997.
- M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85:1501–1510, 2005.
- C.-B. Lee. Estimating the number of change points in exponential families distributions. *Scandinavian Journal of Statistics*, 24:201–210, 1997.
- K. Lin, J. Sharpnack, A. Rinaldo, and R. J. Tibshirani. Approximate recovery in changepoint problems, from ℓ_2 estimation error rates. *arXiv preprint arXiv:1606.06746*, 2016.
- R. McTaggart, G. Daroczi, and C. Leung. Quandl: Api wrapper for quandl.com, 2016. URL <https://CRAN.R-project.org/package=Quandl>. R package version 2.8.0.
- T. Mikosch and C. Stărică. Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics*, 86:378–390, 2004.
- A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557–572, 2004.
- M. Raimondo. Minimax estimation of sharp change points. *Annals of Statistics*, 26:1379–1397, 1998.
- G. Rigai. Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*, 2010.
- E. Ruggieri. A Bayesian approach to detecting change points in climatic records. *International Journal of Climatology*, 33:520–528, 2013.
- A. L. Schroeder and P. Fryzlewicz. Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statistics and Its Interface*, 6:449–461, 2013.

- J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 2012.
- W. Sweldens and P. Schröder. Building your own wavelets at home. In *Wavelets in the Geosciences*, pages 72–107. Springer, 2000.
- A. B. Taylor and R. J. Tibshirani. genlasso: Path algorithm for generalized lasso problems, 2014. URL <https://CRAN.R-project.org/package=genlasso>. R package version 1.3.
- R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42:285–323, 2014.
- UK Land Registry. UK house price index, 2016. URL <http://landregistry.data.gov.uk/app/ukhpi>. [Online; accessed 1-August-2016].
- E. S. Venkatraman. *Consistency results in multiple change-point problems*. PhD thesis, Stanford University, 1992.
- L. Vostrikova. Detection of the disorder in multidimensional random processes. *Soviet Mathematics - Doklady*, 259:270–274, 1981.
- M. P. Wand and M. C. Jones. *Kernel Smoothing*. CRC Press, 1994.
- Y.-C. Yao. Estimating the number of change-points via Schwarz’ criterion. *Statistics and Probability Letters*, 6:181–189, 1988.
- Y.-C. Yao and S. T. Au. Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics*, 51:370–381, 1989.
- A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7:1–38, 2002.
- C. Zou and Lancezhange. nmcd: Non-parametric multiple change-points detection, 2014. URL <https://CRAN.R-project.org/package=nmcd>. R package version 0.3.0.
- C. Zou, G. Yin, L. Feng, and Z. Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of Statistics*, 42:970–1002, 2014.